University of California,
Irvine

The Normative Significance of Volitional Empathy

DISSERTATION

Submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in Philosophy

by

Benjamin Leider

Dissertation Committee:
Assistant Professor Jeffrey Helmreich, Chair
Distinguished Professor Margaret Gilbert
Professor Aaron James

2019

Table of Contents

Acknowledgements

First and foremost, I would like to acknowledge my parents. I know that philosophy was not their first choice for me, but their support and forbearance nonetheless allowed me to complete this doctoral program. I could not have done so without them.

I would also like to thank my committee chair, Professor Jeffrey Helmreich, without whom I likewise could not have completed this program. It was his intervention that allowed me to pass my Portfolio examination, and his enduring patience that allowed me to complete this dissertation despite several time-consuming false starts and wrong turns.

I am also indebted to the members of my committee, Margaret Gilbert and Aaron James, who have been integral to the later stages of my progress in this program. Margaret Gilbert's theory of joint commitment has, in particular, been an inspiration for much of what follows.

I would also like to express my gratitude to the UCI Center for Legal Philosophy, which has generously appointed me Research Scholar for the current year.

Finally, I would be remiss not to acknowledge the late Professor Ted Cohen of the University of Chicago, whose interest in empathy inspired what eventually became the central idea of this dissertation.

Curriculum Vitae

Benjamin Leider

Education

University of California, Irvine — Irvine, CA
PhD - Philosophy — December 2019
MA - Philosophy — December 2014

University of Chicago — Chicago, IL
MA - Humanities — June 2010

Grinnell College — Grinnell, IA
BA with honors - Philosophy — May 2009

Dissertation
Title: *The Normative Significance of Volitional Empathy*
Committee: Jeffrey Helmreich (chair), Margaret Gilbert, Aaron James

AOS: Moral Psychology
AOC: Philosophy of Language, Ethics, Social Ontology

Awards

Machette Foundation Award — June 2016
- *Nominated by philosophy faculty for service to the department*

Teaching

Instructor – University of California, Irvine — Summer 2018
- Philosophy 2 – Puzzles and Paradoxes
- Philosophy 13 – History of Contemporary Philosophy

Guest Instructor – University of California, Irvine — Spring 2016
- Philosophy 144 – Philosophy of Social Phenomena

Teaching Assistant – University of California, Irvine — 2012-2019
- Philosophy 144 – Philosophy of Social Phenomena
- Philosophy 5 – Contemporary Moral Problems: Race and Gender
- Philosophy 4 – Introduction to Ethics (x3)
- Philosophy 5 – Contemporary Moral Problems: Food Ethics
- Philosophy 5 – Contemporary Moral Problems: Poverty
- Philosophy 2 – Puzzles and Paradoxes (x3)
- Philosophy 1 – Introduction to Philosophy (x5)
- Classics 45A – The Gods (x3)
- Classics 45C – Classical Mythology
- Comparative Literature 160 – French Cinema

Graduate Coursework

University of California, Irvine 2011-2015

Ethics and Metaethics

- Philosophy 230 – Virtue Theory 2015 – Winter
- Philosophy 230 – Topics in Ethics: Ethics and Speech Acts 2014 – Spring
- Philosophy 298 – Kant's Ethics 2013 – Winter
- Philosophy 230 – Freedom and Motivation 2012 – Fall

Political Philosophy

- Philosophy 232 – Global Justice 2015 – Spring
- Philosophy 213 – Hobbes and Rousseau 2014 – Winter
- Philosophy 230 – Rights and Obligations 2014 – Winter
- Philosophy 132 – Social Contract 2012 – Fall

Social Philosophy, Philosophy of Law, & Aesthetics

- Philosophy 298 – Independent Study: Philosophy and the Law 2013 – Fall
- Philosophy 244 – Philosophy of Social Phenomena 2012 – Spring
- Classics 200A – Aesthetics of Life 2012 – Winter

History of Philosophy

- Philosophy 213 – Early Modern Theory 2013 – Spring
- Philosophy 215 – History of Analytic Philosophy II 2013 – Spring
- Philosophy 210 – Topics in Ancient Philosophy: *Outlines of Pyrrhonism* 2013 – Winter
- Philosophy 210 – Topics in Ancient Philosophy: Socrates 2012 – Spring
- Philosophy 213 – Kant's First Critique 2011 – Fall

Epistemology

- Philosophy 221 – The Rationality of Belief 2015 – Spring
- Philosophy 221 – The Metaphysics of Knowledge 2014 – Spring
- Philosophy 221 – Skepticism I 2013 – Fall
- Philosophy 201 – Disagreement 2011 – Fall

Philosophy of Mind, Philosophy of Language, & Logic

- Philosophy 250 – Phenomenology 2012 – Spring
- Philosophy 245 – Language and Metametaphysics 2012 – Winter
- Philosophy 205B – Metalogic 2012 – Winter
- Philosophy 205A – Set Theory 2011 – Fall

Latin

- Latin 103 – Livy 2015 – Winter
- Latin 100 – Cicero and Ovid 2014 – Fall

University of Chicago 2009-2010

Logic & Philosophy of Language

- Philosophy 39600 – Intermediate Logic 2010 – Spring
- Philosophy 31414 – Contemporary Analytic Philosophy 2009 – Autumn
- Philosophy 33801 – Theory of Reference 2009 – Autumn

Metaphysics, Epistemology, & Philosophy of Mind

- Philosophy 33000 – Introduction to Metaphysics and Epistemology 2009 – Winter
- Philosophy 38109 – The Philosophy of Wilfrid Sellars 2009 – Winter

Analytic Aesthetics

- Philosophy 31009 – Aesthetics 2010 – Spring
- Philosophy 31210 – Philosophy and Literature 2010 – Winter

Critical Theory

- MAPH 30100 – Foundations of Interpretive Theory 2009 – Autumn

Abstract of the Dissertation

The Normative Significance of Volitional Empathy

by

Benjamin Leider

Doctor of Philosophy in Philosophy

University of California, Irvine, 2019

Assistant Professor Jeffrey Helmreich, Chair

The goal-directed efforts of other agents tend to strike us as to be helped, or at least, as not to be obstructed. We feel called upon in this way almost without any thought, and usually without consulting some general principle of morality, interest, or even courtesy, even where such principles may urge us along the same course. I call this general phenomenon *volitional empathy,* which I take to be the disposition to take what another agent wills as one's own internal reason to will likewise—as a reason to promote, or at least not to obstruct, another agent's goal-directed effort. This dissertation explores volitional empathy's normative significance for regret, punishment, and trust. My first chapter seeks to show that agent regret, understood as moral self-reproach, is rationally required, even though the agent of agent regret is not morally blameworthy for the regretted event. In my second chapter, I argue that punishment, when successful, induces belated volitional empathy in the punishment subject, and thereby prompts the punishment subject's regret for the punished violation. This effect of punishment raises the prospect that volitional empathy might underwrite interpersonal norms by giving one agent an internal reason under the control of a second agent, namely, the agent whose effort underwrites the internal reason of the first. In my third chapter, I use this idea to explain the normativity of trust: I propose that the trustor's reliance on the trustee involves the volitional empathy of the trustee for the trustor. The trustor can enforce this reliance by compelling either the trustee's performance, or, following betrayal, by compelling him to repair the harm he caused. In both cases, the trustor pressures the trustee with his own internal reason, which, being his own, he cannot but take as motivating. The last chapter more closely considers rebukes of the betrayal of trust. Since rebukes

occur after the damage of betrayal is done, it is not obvious why they are reasonable. I show that they are reasonable by explaining how, by inducing regret, they allow the trustor to compel the trustee to repair the harm he caused.

Introduction

You reach for a book perched on a high shelf deep in the stacks. You brush the spine with your fingertips, but the book remains beyond your grasp. Just feet away, hunched over a book of my own, there I stand, unknown to you, unlikely to see you ever again—I'm visiting from out of town. For a moment, my eyes drift above my book, and I see you straining to reach one of the upper shelves—for you, a tragic inch too high, but not for me. Should I help you out?

You're in the mailroom of your apartment building, separating the bills and notices from the PennySaver, the catalogs, and the alumni magazines. You catch sight of a person standing outside the building's glass front door just beyond the mailroom entrance. She's a resident—a casual acquaintance—and she carries in each arm a large bag of groceries. Preparing to unlock the door, she fumbles with her keyring, heroically refusing to set the bags down. One of them starts to slip. She swiftly shifts to save it, but her hold on the other bag becomes more precarious. Please, just put down the bags, you think at her to yourself. Should you help her out?

These are not the questions I investigate, but rather, those whose common answer I take for granted. And that answer is: yes, of course! Of course I should help you fetch the book from the high shelf. Of course you should open the door for the hapless laden resident. And of course, this answer assumes that other things are equal. If you're in crutches, perhaps the reasonable extent of the help you render is to silently wish the best to the woman outside. If I see the edges of a suspicious geometric tattoo peeking out from the receding sleeve of your extended right arm, which turns out to be reaching for *Mein Kampf* in a way, now that I think about it, strangely reminiscent of how the Germans reached for Hitler, then such considerations might reasonably give me pause.

What I am pointing towards here is that when a goal-directed effort is salient to us, it tends to strike us as something to be helped along, or at least not to be interfered with. I say it "strikes us" this way because this normative appearance is not mediated by an inference from a higher principle or an ulterior motive; rather, it's just the way that a goal-directed effort looks—the way a *volition* looks, to use the term I prefer in the chapters that follow. We are just disposed to find the volitions of others as guiding our practical reasoning towards helping or not interfering. That is to say, we are disposed to treat the volitions of others as *pro tanto* practical reasons, whether or not they really are—whether or not they *ought* to guide our practical reasoning in the way they actually do. In this way, the volitions of others behave like our own ends. Perhaps they might even be counted among our own ends, insofar as we might acquire ends from different sources, and insofar as we cannot simply acquire or discard ends at a whim. I have therefore decided to call these reason-like things *internal reasons*, since they seem to fit what Bernard Williams describes under that name, though without committing to reasons internalism—or to reasons externalism, for that matter.[1]

Volitional empathy is my name for this disposition to take the volitions of others as internal reasons of one's own. Volitional empathy is volitional twice over: first, because it is a responsiveness to the volitions of others; and second, because it guides us towards having volitions of our own. It is empathy a broad sense—or analogous to empathy, if we wish to reserve that term for emotions—because in treating the volitions of others as reasons, we are prompted to have volitions with the same content: I take your volition to retrieve the book as an internal reason of my own to retrieve the book for you. Unlike pure emotional contagion, however,

[1] Williams 1982a. I distinguish under the name *volitional reasons* those internal reasons that we acquire from the volitions of other agents.

volitional empathy does not bypass our reasoning, but rather submits an input to it. After all, the input may be outweighed by the balance of other considerations, as I have mentioned.

The overarching goal of this dissertation is to explore volitional empathy's normative consequences. Here, by 'normative', I mean normative in a weak, agent-relative sense that internal reasons can, I hope, be uncontroversially normative: guiding of an agent's practical reasoning in the non-compulsive, inferentially sensitive way that reasons ought to guide it. This guidance is not purely causal as it involves the following idealization: the agent deliberates in an instrumentally rational way with adequate relevant information—that is, in a way not liable to undermine her own ends. I do assume that this sense of normativity is *largely* predictive of agents' practical reasoning, since I assume that it is possible to impute ends to most agents such that they are largely instrumentally rational, and since I assume that the best interpretation of their ends is the one most charitable in this way.

In my first chapter, which directly considers agent regret but speaks to regret more generally, volitional empathy does not explicitly arise. I do, however, make ample use of internal normativity to show that an agent's *own* ends can make it appropriate for him to experience regret when he is morally blameless, and in just the same way they do when he is morally blameworthy. In addition, I conceive of regret as a kind of volition, and consequently, something that volitional empathy can prompt. Regret figures prominently in my second chapter, where I take advantage of its volitional character to argue that punishment, when successful, induces the punishment subject to feel regret by prompting his belated volitional empathy. The idea that punishment can, and perhaps should, induce volitional empathy, raises the possibility that volitional empathy can underwrite interpersonal norms. After all, if one can prompt an agent to

experience volitional empathy, then one can make doing what one expects of him to be "right" by his own lights, even against his own wishes. I develop this idea in my third chapter, where I propose that trust involves reliance required by the volitional empathy of the trustee for the trustor, and enforceable by the trustor in the much the same way that rights are: the trustor can demand that the trustee keep her trust, and rebuke him for breaking it, and in this way restore either his motivation not to betray her, or his motivation to repair the harm she suffers after his betrayal. This chapter focuses on how volitional empathy can underwrite the normativity of trust, and touches on demands and rebukes only briefly. My last chapter considers demands and rebukes more closely. It attends to rebukes in particular, since it is unclear why one should bother to rebuke someone who has already betrayed one's trust. My answer recalls the discussion of regret from the second chapter, as I propose that rebukes prompt regret, while regret prompts repair, both of the harm the trustor suffered because her reliance was disappointed, and of the harm to her relationship with the trustee on account of his betrayal disconfirming her belief that he is trustworthy.

Chapter 1: Agent Regret and Practical Reason

Suppose on a trip out of town you spot an old friend you haven't seen in decades. You are positively thrilled, and in your eagerness you approach him and call out, "Jack, is that you?" Unfortunately, Jack over the years has developed a serious heart condition. Startled, he clutches his chest and collapses. Or suppose you're an electrician working with your partner on the wiring in a house. She calls down to you, "I think I've fixed it! Turn the circuit breakers back on—I want to see if everything works now." Moments later, she realizes she's dropped a tool onto an exposed wire. She reaches for it just as you hit the circuit breaker switches, and is electrocuted. Or consider Bernard Williams's canonical example: while driving down a street, a child darts into the path of your truck. Although you are driving carefully and at an appropriate speed, you run the child over before you can react, killing him.

In all of these cases the agent does nothing wrong. It even seems appropriate to comfort the agent as if he were morally in the same position with respect to the misfortune he caused as would be an innocent spectator. And yet, Williams maintains, "There is something special about [the agent's] relation to this happening, something which cannot merely be eliminated by the consideration that it was not his fault"[2]. Williams calls the agent's self-reproachful attitude towards her special relation to what happened *agent regret,* and he intimates throughout his article that this attitude has a moral character. I wish to consider the following question: is agent regret, understood as self-reproach of a moral kind, in fact an appropriate response to causing harm through no fault of one's own?

[2] 1982b, 28

Harry Frankfurt and David Sussman do not hold that such a response is appropriate. They agree that a negative attitude of *some* kind is appropriate, but it is non-moral in character. Julie Tannenbaum, Jeffrey Helmreich, and Connie Rosati come down on the other side of the question, holding that moral self-reproach is in fact appropriate on the agent's part. R. Jay Wallace and Joseph Raz disagree over whether the agent's response is of a different kind that the spectator's response; the truth of the matter will determine whether a rationalization of the agent's response must also rationalize the spectator's response.

In this paper, I consider and critique these positions. I then develop and argue for my own answer: agent regret understood as moral self-reproach is appropriate because it is rationally required, and it is rationally required because innocent moral agents ought to reproach themselves in broadly the same way that guilty moral agents ought to reproach themselves. That is to say, innocent moral agents, like guilty moral agents, are rationally required to have past-oriented volitions that can no longer be carried out because it is now too late to do so.

This paper has two parts. The first discusses the positions of Harry Frankfurt, David Sussman, Julie Tannenbaum, Jeffrey Helmreich, Joseph Raz, R. Jay Wallace, and Connie Rosati on whether moral self-reproach is appropriate for an innocent agent who brings about harm. I invite readers mainly interested in my positive account to skip to the second part of this paper, beginning on page 30. This part consists of the last three sections, beginning with section seven, where I lay out certain preliminary assumptions of mine about the nature of instrumental rationality. In section eight I distinguish temporal features of volitions that allow regret to be a volitional phenomenon despite being a past-oriented attitude. I bring everything together in section nine, where given the volitional coherence principle discussed in section seven and the

temporal features of volitions discussed in section eight, I conclude that it is in fact appropriate—even rationally required—for an innocent agent to experience moral self-reproach.

Part 1: The Agent Regret Literature

I. Frankfurt

For Harry Frankfurt, when an agent is innocent, it cannot be appropriate for him to experience agent regret understood as *moral* self-reproach. Like Williams and the other authors I consider, Frankfurt accepts that an agent's negative reaction will normally follow blamelessly causing harm. He concludes that it is in fact appropriate for the blameless agent to reproach himself, though only in a non-moral way. Frankfurt characterizes this non-moral species of self-reproach as *embarrassment* for "having failed," for being "deficient."[3] Embarrassment appropriately intensifies by degrees into *shame* and eventually *horror* in proportion to the severity of the action's consequences. It may even be appropriate for the agent to think of himself as a "*poisonous creature,* who cannot avoid doing dreadful harm," and even to conclude "*that the world would be better off without him,*" while at the same time "acknowledging no *moral* responsibility at all for being so toxic," as that acknowledgement would *not* be appropriate.[4]

There is much to be said for Frankfurt's account. It takes seriously the fact that people reproach themselves for causing harm even if they consider themselves morally innocent. In addition, it plausibly distinguishes self-reproach—the negative attitude of the agent—from the negative attitude of spectators: although spectators may find the outcome of the agent's action

[3] Frankfurt 2008, 10–11
[4] Frankfurt 2008, 13

unfortunate, they will not attribute that outcome to their own "deficiency or an inadequacy," and will have no reason to reproach the agent for his, either [5]. [6]

Although I won't dispute the claim that agent regret concerns deficiency or failure, I'm not convinced that embarrassment or shame for this deficiency is the most complete or accurate way to characterize the feeling occasioned by blamelessly harming another person. For one, embarrassment can be understood rather light-heartedly, as when someone stumbles or drops something. Frankfurt has foreseen this worry. He answers:

> Perhaps it may seem that embarrassment is not a sufficiently penetrating or portentous emotion to be suitable as a response to having killed someone or having caused someone to die. It may strike us as too shallow to reflect at all adequately a person's recognition that he has brought about an immeasurable and irreparable harm. In fact, however, a feeling of embarrassment may be both deep and shattering. It need not be shallow or inconsequential. After all, embarrassment is closely related to *shame*; and feelings of shame may be quite devastating.[7]

I reply, in turn, that the problem with characterizing agent regret as 'embarrassment' is not the intensity, magnitude, or severity of embarrassment as a mental state, but rather: (1) that embarrassment and agent regret are *about* different things, and (2) that embarrassment and agent regret are phenomenally distinct—they simply feel different. We can discover both differences using Frankfurt's own examples: we may see how (1) bears out by considering his example of "trivial" embarrassment, and we may see how (2) bears out by considering his example of severe embarrassment.

---

[5] Frankfurt 2008, 12
[6] Unless he could have intervened, in which case he might also experience self-reproach to some degree.
[7] Frankfurt 2008, 10

First let us consider Frankfurt's "trivial" example:

> To take a trivial example… [the feelings of embarrassment] may be the very feelings that would be expected of someone who, in the midst of a formal dinner party, emits a loud and grossly offensive belch. Perhaps he really could not help himself. Let us assume, at any rate, that the belch was not voluntary, and that it was truly uncontrollable…. Nevertheless, he will quite naturally—and indeed, quite appropriately—chastise himself; and he will chastise himself *precisely* for *having failed* to notice that it was coming, and for *having failed* to suppress it.[8]

But now suppose that the moment before the belch escapes him, a waiter drops a large tray of glasses filled with champagne. There is a spectacular crash as the glasses explode against marble floor, diverting the attention of the other guests and completely concealing the belch. It will still be the case that our now-fortunate guest failed to notice his belch coming, and it will still be the case that he failed to suppress it. And yet he will not be embarrassed but relieved. This is because he would not be embarrassed *that he belched at a formal dinner party,* but *that the other guests—the president of the country club, the senator, the boss, and the woman he was chatting up—all* saw *him belch at a formal dinner party*. Paradigmatic of embarrassment is that it concerns *other people finding out*. True agent regret, however, cannot be defeated in this way: the truck driver will be riven with guilt-like feelings of self-reproach even if the street was entirely deserted but for the child, whose lifeless body rolls into an open manhole and washes out into the middle of the ocean, never to be discovered. This indefeasibility is not a result of the severity of the harm in the truck driver example, either. Suppose I slip on ice outside of an opening door and use it to catch myself, at the same time forcing it closed and—unbeknownst to me—painfully tweaking the hand

[8] Frankfurt 2008, 10–11

of the person trying to exit. There are no witnesses. I discover later that this person was injured by my action and that she blamed the wind, which was really gusting that day. It is quite plausible that I reproach myself, and yet if I do, it will *not* be on account of people finding out, as in this case they don't.

Now let's consider Frankfurt's example of severe embarrassment:

> Let us suppose… that a person is the carrier of a highly contagious and dreadful disease. Mere proximity to this person, even without any more intimate contact, is sufficient to lead to infection with a severely debilitating and often fatal illness. Let us say, moreover, that this person came to be a carrier of the disease through no fault of his or her own. It was entirely inadvertent that the person became a terrible threat to everyone around. It was just a matter of bad luck that the world became worse because of this particular person's misfortune in acquiring the uncontrollable tendency to spread illness and death.[9]

I grant that the agent will in this case feel terrible even if we stipulate that no one discovers his role in the spreading of his illness. But this is not obviously an example of embarrassment at all, severe or otherwise. If anything, this is an example of blamelessly inflicted harm, much like the harm blamelessly inflicted by Williams's truck driver. For that reason, it merely raises the question of whether agent regret is really embarrassment, rather than showing it to be so. Contrast this with a more obvious and central case of severe embarrassment. Suppose one is using PowerPoint slides to deliver a talk on a dry subject to a packed lecture hall. One advances to the next slide, only to discover that it has been replaced with a salacious photograph of one engaged in sexual relations with one's spouse. I submit that one would find oneself severely embarrassed. I

[9] Frankfurt 2008, 13

submit further that one would be embarrassed before it even occurred to one how the photograph got into the slideshow, or how it was taken (we'll stipulate that one doesn't arrange for such photographs), or whether one was negligent in not inspecting the marital bedroom for hidden cameras, or anything else at all that might be the content of self-reproach. I submit, finally, that in this surely conceptually central case of severe embarrassment, the horror one experiences at the moment of exposure would feel very different from the horror one might experience upon learning that one is a highly infectious carrier of some dreadful disease.

Frankfurt is driven to his view because he reasons that only being guilty can make *moral* self-reproach appropriate. Therefore, without guilt, something other than moral self-reproach—such as embarrassment—must be appropriate. The trouble Frankfurt has with cashing out this alternative suggests that Williams may be right about the moral character of innocently causing harm, a character that *feeling terrible* and *feeling embarrassed* do not acknowledge, much less explain.

II. Sussman

Sussman agrees with Frankfurt that agent regret is appropriate but non-moral. It is appropriate for Sussman because it is rational, and it is rational because it consists in part of desiring to make "amends" to a person one has blamelessly harmed while sharing with her a temporary "moral state of nature" where each person is morally entitled to kill the other.[10] Sussman asks us to consider the following case, somewhat altered from Williams's example:

> In place of the child, consider instead an adult pedestrian blown into the path of a truck by a freak gust of wind and badly injured (rather than killed) by that truck. I

[10] Sussman 2018, 792, 804–5

> assume that, like the truck driver, the pedestrian is without fault; she's sober, watching where she's walking, wearing sensible shoes, and so on. What's important to see here is that as he careens toward this unlucky person, the truck driver comes to stand toward her as a kind of innocent threat.[11]

In an example reminiscent of Judith Thomson, and with the same conclusion, Sussman asserts that the pedestrian would be morally entitled to use her disintegrator gun to obliterate the oncoming truck along with its driver whom she correctly believes to be innocent. But as a consequence, "as the pedestrian raises her disintegrator to vaporize the approaching truck, so too may the truck driver draw his own disintegrator in an effort to preempt her attempts at what he knows to be completely justified self-defense."[12] As it happens, neither the pedestrian nor the driver are packing heat that day, and fortunately, the pedestrian is not killed but only injured. While better than the outcomes in which either party is killed, there is now an awkwardness, to say the least, between the truck driver and the injured pedestrian.

Sussman feels a need in this case for the truck driver to issue a "quasi-apology," where quasi-apologies, or at least the disposition to make them, is a component of agent regret. A quasi-apology is a kind of fictional apology—an apology in form, but under the mutually understood false pretense that the truck driver is blameworthy and that the pedestrian is morally free to reject it and consequently to decline to associate with him. Sussman compares this exchange with "fighting" with his wealthier father over the privilege of paying a restaurant bill—a privilege that his father always "wins." Both Sussman and his father understand that the latter *should* pay, as he is richer, but also that if he does not, then Sussman *should not* protest, having himself demanded

[11] 2018, 802
[12] 2018, 803

to pay. As in Sussman's ritual with his father, both the truck driver and the pedestrian understand that the truck driver is not at fault and that the pedestrian *should* "forgive" him—that is, "quasi-forgive" him—but if she declines, then the truck driver *should not* protest.

Sussman variously describes the purpose of these quasi-actions as effecting a "joint affirmation" of the cessation of the moral state of nature "by which their return to a way of living together is acknowledge and ratified"; as facilitating a "return to something like the space of social possibilities that existed before the conflict"; and as meeting a "need to recognize that we have done (blameless) wrongs" and "to disown those wrongs while nevertheless recognizing them as blameless," for "after such a crisis, these people cannot just go on as they had before."[13] The idea seems to be that there is a social relationship, or at least the potential for one, that is damaged when agents enter and emerge from a moral state of nature, even though everyone emerges with morally clean hands.

Sussman here uncovers a fascinating phenomenon by drawing attention to the strange normative position of the driver vis-à-vis the pedestrian. On the one hand, by quasi-apologizing, the truck driver affirms the pedestrian's privilege[14] to resume or not the social status quo ante. On the other hand, by *expecting* to be quasi-forgiven (precisely as someone who *actually* apologizes cannot expect), the driver presupposes a claim[15] against the pedestrian failing to resume the status quo ante, which is precisely to *reject* the pedestrian's privilege to resume the status quo ante or not. This normative contradiction can be resolved by distinguishing two logically independent

[13] 2018, 804–5
[14] Strictly speaking, this is probably a higher-order privilege—an Hohfeldian *power*—to resume or not the claims and duties constituting the social relationship in the status quo ante.
[15] Strictly, an Hohfeldian *immunity*.

kinds of normative relations: moral relations and nonmoral social relations, such as social relations conceived in terms of Margaret Gilbert's theory of joint commitment.

On such an understanding, in virtue of being morally blameless, the truck driver has a *moral* claim against the pedestrian declining to resume the social status quo ante. But as the truck bore down upon the pedestrian, both people—had they a moment to reflect on this—would recognize the pedestrian's privilege to kill the truck driver in order to save herself. This recognition arguably constitutes a mutual abrogation—a "joint rescission," in Gilbert's terms—of the nonmoral claims and duties following from the joint actions, joint beliefs, joint decisions, and other collective intentionality phenomena that constitute the social relationship between the truck driver and the pedestrian.[16] These claims and duties, according to Gilbert, are those that are necessary for people doing things *together* to mutually conform to their shared ends [17]. Needless to say, refraining from killing each other is necessary for virtually all[18] shared ends, and thus constitutes a nonmoral duty (independent of any similar moral duties) owed by each participant to the others. This very duty is what the pedestrian and the driver jointly rescind. The social relationships constituted by the shared ends depending on that duty are therefore suspended until they are restored by a process of "joint commitment" of their participants to resume them.[19] And indeed, by a convergence of language reflecting perhaps a similar convergence in thought,

[16] 2018, 165, 192–93
[17] 2018, 169
[18] But not quite all, e.g., Josephus's suicide pact with the other zealots to be carried out by mutual execution.
[19] That is to say, the participants create claims and correlative duties upon each other constraining them to φ as a body (that is, to "emulate, by virtue of their several actions… a single non-collective body" in φ-ing) 2015a, 119.

Sussman speaks of a need after the crisis for a "joint affirmation by which their return to a way of living together is acknowledged and ratified."[20]

Sussman's account is most illuminating where it discusses the peculiar normative situation of the agent vis-à-vis the victim. It rationalizes agent regret as the attitude that prompts the agent to recognize this situation, but as we shall also see when discussing Tannenbaum, it fares less well in rationalizing agent regret's affective character. Unlike Tannenbaum, who explicitly brackets agent regret's affective character, Sussman seems to assume this explanatory burden when he criticizes Williams for observing without explaining the remorse-like character of agent regret (791-2). But it is not clear why the truck driver's restoration of normal social life with the pedestrian *must* have the affective valence of an apology, or, for that matter, why it must have the form of an apology at all, rather than the form of a frank conversation about the normative facts on the ground (whether there is fault, whether there is harm, and whether the pedestrian wishes to go on as before). At the very least, the necessity of the pretense of apology, rather than a frank conversation, is no clearer than the necessity of the pretense of fighting over the dinner check with one's father rather than gratefully and humbly accepting the meal.

Finally, Sussman's account doesn't explain why agent regret is appropriate when quasi-forgiveness is impossible. Recall that Sussman changes Williams's example so that the person hit by the truck is an adult blown into the street by the wind, and who is badly injured but not killed. Sussman wishes to bracket "concerns that depend on the special moral status of children" and also, presumably, to silence any suspicion that the adult pedestrian might herself be at fault.[21] But

[20] 2018, 805
[21] Sussman 2018, 802

altering the example to preserve the pedestrian's life betrays a lack of generality in Sussman's rationalization of agent regret. Suppose that the pedestrian had been a vagrant with no friends, co-workers, or living relations—in short, someone who would not be missed, at least not by anyone other than her killer. The driver cannot seek quasi-forgiveness from someone who no longer exists, nor from any people who might quasi-forgive him on the vagrant's behalf, or on their own behalf as people—such as a friends or relatives—who are themselves indirectly harmed. In this case, it is impossible to restore normal social life with the victim or with anyone else by quasi-apology, and yet this is far from a peripheral case of agent regret. Indeed, in Williams's original example the victim is killed. If the rationality of agent regret depends on the rationality of restoring normal social life, then Sussman does not explain the rationality of agent regret in the paradigmatic case of the phenomenon.

III. Tannenbaum

Perhaps Frankfurt feels that only moral blameworthiness can make moral self-reproach appropriate because he assumes a representational view of the appropriateness of a state like moral self-reproach, a view of the sort advanced by Julie Tannenbaum. On this view, moral self-reproach for an action is appropriate insofar as the self-reproachful attitude somehow correctly represents to oneself the action's wrongness. Given this representational story, one must hold either that moral self-reproach is inappropriate because the agent clearly isn't morally blameworthy; or that moral self-reproach is appropriate because the agent really is morally blameworthy, where the appearance to the contrary is explained away. If Frankfurt in fact relies on such a representational view, he has opted for the first alternative. Tannenbaum, as we shall

see, opts for the second. Part of my proposal is that, at least as far as agent regret is concerned, the representational view is mistaken.

In "Emotional expressions of moral value," Tannenbaum examines agent regret in Williams's paradigmatic case of the truck driver. She takes this instance of agent regret to be a "moral feeling"—in this case, a feeling associated with a special negative moral self-evaluation for an action. Since the feeling is associated with the negative moral evaluation, Tannenbaum thinks that it is rational insofar as the evaluation "correctly represents the world."[22] A negative moral evaluation correctly represents the world, in turn, when the action evaluated satisfies at least one of the three following conditions:

(1) The action is not chosen for a morally valuable end.

(2) The action does not realize the agent's end or sub-ends.

(3) The agent is acting for the wrong end.

Failing to satisfy (1) or (3) justifies the evaluation that the agent violated his obligations. Failing to satisfy (2) justifies the evaluation that the agent failed to realize his moral ends or associated sub-ends, such as not killing children. It is this latter evaluation that Tannenbaum associates with agent regret.[23]

Tannenbaum's argument offers a valuable interpretation of the action in examples like Williams's, an interpretation that contributes to rationalizing both the negative moral self-evaluation on the agent's part, and the judgment of the agent's blamelessness on the part of everyone else. And in fairness to Tannenbaum, this is the precisely the declared goal of her

[22] Tannenbaum 2007, 44
[23] 2007, 45

paper—to show that agent regret meets a "necessary condition for being rational" because "the negative moral evaluation associated with agent-regret correctly represents the action as morally inadequate."[24] In a footnote to the quoted text, however, Tannenbaum seems to formulate her goal in stronger terms, ironically by way of a disclaimer: "In this paper," she writes, "I do not argue for or against the claim that the driver's feelings are rationally (or morally) required, but rather that they are justified. If a feeling is required, then lacking the feeling would be unjustified."[25]

If we are to understand that her goal is merely to show that there is a way in which agent regret is *not unjustified,* or that there is a feature of agent regret (say, the representational accuracy of its thought content) that *contributes to its justification,* then we may without further ado look beyond her account as we investigate how agent regret is positively justified. On the other hand, if we understand Tannenbaum as also seeking to establish positive justification of agent regret, then her argument, I suggest, fails to account for the guilt-like feeling that Williams first observed in his case of the truck driver. Recall that the goal in the present paper is to justify the specific experience of agent regret when it is inappropriate for anyone to blame the agent but himself. What is missing from Tannenbaum's story is an account of why it is *agent regret* that is justified instead of some other feeling associated with the relevant negative moral self-evaluation.

Her argument, in other words, would justify too much: it would show not only that agent regret is justified, but that *any* feeling is justified if it happens to be associated with an accurate negative moral self-evaluation, *whether that feeling is the guilt-like self-reproach of agent regret or not.* Perhaps one associates with negative moral self-evaluation the feeling of equanimity, or

[24] 2007, 44–45
[25] 2007, 45n9

cynical resignation, or even glee, where some possibilities are surely less justified than others. So long as this feeling is associated with the negative moral evaluation she describes, and so long as the evaluation itself "correctly represents the world," or in this case the action, the feeling counts as justified. These two conditions, however, seem quite insufficient for rationality because agent regret is not just its thought content but also an experience with phenomenally distinct affect: it *feels* a certain way. A state involving Tannenbaum's evaluative thought content but accompanied by glee would not only not be agent regret, it wouldn't be rational. And lest we be tempted to deny the very possibility that thought content can be associated with incongruous affect, John Deigh has offered the counterexample of irrational fear, which can exist without the appropriate evaluative judgment—the judgment that the fear object is a "potential source or agent of some bad effect," as in phobic episodes (e.g., involving a garter snake one knows to be harmless) and instinctive reactions (e.g., to loud noises).[26]

IV. Helmreich

Unlike Tannenbaum, Helmreich holds that it may indeed be appropriate for an agent to have a self-critical reaction to causing harm *even if she is unequivocally blameless*. "[H]arming someone in any way," he writes, "including legally and morally innocent ways, merits a self-critical stance." His argument is thus:

(i) "Moral agents are deeply invested in not harming others."[27]

(ii) "Investment against harm grounds self-criticism" should they harm others.[28]

[26] Deigh 1994, 836–37
[27] Helmreich 2011, 585
[28] 2011, 586

(iii) Therefore, moral agents are grounded in their self-criticism should they harm others.

Helmreich justifies (ii), that "investment against harm grounds self-criticism,"[29] by observing that when agents are invested in a goal, they take a self-critical stance towards accidentally subverting it. In addition, the intensity of self-criticism scales with the degree of investment. To use Helmreich's example, one would surely take a highly self-critical stance towards accidentally destroying a house one was building. This stance is a self-critical one (as opposed to a stance of some other kind, presumably) because self-criticism is simply how people experience the "tension between what one actually did and what one was deeply invested in doing."[30]

Helmreich may be saying one of two things here: (a) that a self-critical orientation towards failure is a *universal* response of agents, or (b) that a self-critical orientation towards failure is a *common* response of agents. In fairness to Helmreich, (b) is enough to settle the titular question of his article, "Does sorry incriminate." For regardless of whether it is rational to do so, if people are prone to apologize even when conscious of their own innocence, then apologies do not reliably correlate with wrongdoing, and therefore do not incriminate any more than polygraphs do. That said, Helmreich is clearly reaching for something more ambitious than (b)—after all, he writes that the innocent harm "merits" a self-critical stance—but his argument does not support (a). People no doubt *tend* to be self-critically oriented towards failure, and if I were to observe someone apparently indifferent to it, then I would likely suspect the person's commitment to succeeding. But it also seems possible that my suspicion is mistaken. Helmreich allows as much

[29] 2011, 586
[30] 2011, 586–87

by granting the conceptual coherence of a "legalistic moral agent" who somehow fulfills her obligations without the sort of investment that produces a self-critical stance towards failure, but he doubts that legalistic agents would tend to fulfill their moral obligations successfully.[31]

This doubt assumes that investment, and thus the self-critical stance it entails, contributes to success. However, a case suggested by Helmreich's own example involving competitive sports may argue otherwise. While it is *common* to assume a self-critical stance towards, say, missing the basket when shooting a basketball, this stance can interfere both with one's subsequent ability to play and with one's ability to have made the failed shot to begin with. For the stance of self-criticism is experienced not only subsequent but also prior to the attempt, when it takes the form of performance anxiety. Moderating this stance for the sake of performing well is especially critical when the stakes are high—when one's investment is great—and no stakes are higher than moral stakes. Like athletes who take their minds of winning, moral agents may even *divest* themselves of the very goals they seek to promote to the extent that investment interferes with success, even if the optimal level of investment underrepresents the moral urgency of the goals in question.

V. Raz and Wallace

An undisputed feature of agent regret is what Joseph Raz calls its "self-reflexive" character. That is to say, agent regret is a *de se* attitude: *x* regrets that *x* φ'd, where the identity of the referents of the first and second occurrences of 'x' is known to x and is necessary for regret of this kind. For example, Oedipus's initial regret that *someone* killed the previous king of Thebes is not

[31] 2011, 589–90

agent regret, even though it will turn out that he is the person in question. But after discovering the identity of the perpetrator, his regret that *he* killed the previous king of Thebes—*he* rather than someone else—is agent regret. The identity of the regret subject with the subject of the regret's content is significant for Raz because it makes the attitude "particularly poignant due to its being, in part, about the person one is or was, as manifested on that occasion." Raz elaborates that "it is poignant in being, not regret that there is such a person, but that I am such a person."[32]

I mention Raz's discussion of self-reflexivity because my rationalization of regret concerns agent regret only, as opposed to non-self-reflexive "spectator regret" of the sort an uninvolved witness might experience upon seeing the truck driver kill the child. As we shall see, my argument derives the rationality of agent regret from agent-relative reasons following from instrumental rationality. This avenue of justification is not open to spectator regret. Consequently, it matters to my account that agent regret is not merely a case of spectator regret, *pace* R. Jay Wallace, who undermines the strategy Raz uses to distinguish agent and spectator regret from each other. Wallace writes:

> The uninvolved observer or third party, after all, can perfectly well appreciate that the unfortunate event occurred, and also that it is unfortunate…. So what proprietary form of conscious awareness might be taken to be available only to the agent? Is it just the indexical thought that it was *my* agency that gave rise to the unfortunate event? It is true that an emotional reaction that constitutively involved an indexical thought of this kind would not even be available to those who were not themselves caught up as agents in the regrettable events…. But it isn't obvious that emotions of retrospective assessment that involve such indexical thoughts really constitute an interesting natural kind of psychological

[32] Raz 2012, 141

> phenomenon. Nor is it clear why it should be important to insist that agents such as the lorry driver rightly feel emotions of this kind.... What exactly would be deficient in the attitudes of a lorry driver who felt profound and persistent regret about the fact that the child died, but only of an impersonal kind that was not tied constitutively to the thought that the death resulted from his agency in particular?[33]

Raz can cash out the "indexical thought" that it was *I* who did such and such as the thought that I am the *sort of person* who does such and such. But this analysis does not distinguish agent regret from third-party attitudes: a mother disappointed in her son might well regret that *he* is the sort of person who does such and such.

I must therefore agree with Wallace that there is little promise in strategies like Raz's for distinguishing agent from spectator regret. But there is another way of establishing the distinctness of agent regret, a way following from a feature of agent regret—or rather of what Wallace considers regret in general—as recognized by Wallace himself. He observes that all regret involves an "orectic attitude" not supplied by a mere evaluative judgment, such as the judgment that things should have been otherwise. Wallace initially describes this attitude as being supplied by a "wish" or "preference" that things were otherwise, over and above a mere evaluative judgment that they *should* have been otherwise.[34] But Wallace seems to be reaching for something stronger than wishing or preferring, as he describes regret as closely analogous to *intention,* and specifically distinguishes "desires, construed as attraction or repulsion, and intentions," where one can desire something without actually intending it. Furthermore, "[i]ntentions represent commitments of the kind that bring practical deliberation to a provisional

[33] Wallace 2017, 36
[34] 2017, 49

conclusion, whereas desires are states that precede deliberation and provide a potential for choice."[35]

I gather, then, that by 'orectic attitude', Wallace means a sort of striving stronger than mere desire—strong enough to issue in action if only there were an opportunity, as there often is in the case of garden-variety intentions. I entirely agree that regret has this conative character, as I will call it, and I am very sympathetic to the analogy with intention. My own account will assimilate regret to a species of volition, where I say 'volition' in part[36] to avoid the theoretical commitments of the term 'intention'.

The close analogy of regret with intention is important to Wallace because it allows regret to have intention's orectic character. Indeed, Wallace seems inclined to consider regret a species of intention but for the fact that regrets are about the unchangeable past, while "intentions typically take as their objects prospective actions that represent real options for the agent at the time when they are formed."[37] But if regret and intention are really so similar—if they are really different only in respect of the "reality" of their objects as options for the agent—then as I shall presently show, it is difficult to resist the consequence that this account of regret in general is in fact an account of *agent* regret. As for spectator regret, it must be something fundamentally different, perhaps in being desire-like rather than intention-like, since Wallace's analysis seems to exclude it.

---

[35] 2017, 55

[36] This is not my only motive for choosing the term 'volition', however. 'Volition' is more native to psychological discourse while 'intention' is more native to practical reasoning discourse. I will argue that regret is a volition about the past, which, while strange-sounding, is less paradoxical than an intention about the past. That is to say, while there is certainly nothing to be *done* about the past, perhaps there is yet something to be *willed* about it.

[37] 2017, 56

The reason Wallace's account of regret in general is in fact an account of agent regret is that there is reason to think that *intentions* are always self-reflexive. Consider how the locutions for intention statements nearly always follow the word 'intend' or 'intention' with an infinite clause. For instance, I intend to convince you that Wallace is mistaken. I hope it is your intention to hear me out on this matter. This self-reflexivity is more explicit when we restate these intentions in terms of *that* clauses: I intend *that I* will convince you that Wallace is mistaken. You intend *that you* will hear me out on this matter. While clumsy, these constructions reveal the subject of the content of the intention to be the very person to whom the intention itself belongs—they reveal that these intentions are self-reflexive.

Gilbert Harman has argued that *all* intentions are self-referential in a way that makes them self-reflexive: intending to φ is always intending to φ *by that very intention to* ϕ. This feature of intention is necessary, for among other reasons, for the right actions to count as intentional as opposed to unintentional actions—intentional actions being those that result non-deviantly from their agents' intentions. To adapt Harman's example, if Mabel could merely intend that Ted be killed (rather than intending that because of that very intention, she kill Ted), then it would be difficult to make sense of how she kills him unintentionally when she accidentally runs him over while driving to his house.[38] Against this point, Luca Ferrero claims that a further argument is needed to show that an intended state of affairs (the object of an intention *that*, to use a construction grammatically permitting non-reflexive intentions) must be the intention subject's own action.[39] But it is hard to see how an intended state of affairs could possibly result non-deviantly from one's intention without it being true that one *saw to it* that the state of affairs

[38] 1976, 441–45
[39] 2013, 80

obtains: there is something paradoxical about the following conjunction: I intended *that* φ, φ results non-deviantly from my intention, but I did not see *to it* that φ.

VI. Rosati

In spite of the disagreement voiced above, I am very sympathetic to Wallace's analogizing of regret with intention, an analogy I will echo in my own account by considering regret to be a species of volition. I am also sympathetic to Connie Rosati's account of the how the cognitive (as opposed to the affective) component of nonmoral agent regret can be "well-grounded"—that is, how the "normative assessments it involves are true or warranted."[40] Rosati's account is shaped by the intuition that "a good life is a life without cause for regret"—which she takes to mean that regrets are somehow grounded by reference to a *conception of the good*.[41] Her account is also shaped by her position that one's good can change over time. She derives this position from the fact that regrets are subject to reevaluation, rather than forever striking us as irretrievable errors that permanently reduce the value of our lives.[42]

Rosati introduces the notion of an *effective conception of the good*, which we individually develop for ourselves in order to organize our desires into a "coherent, stable, and attractive set of aims" that suit our changing "abilities, temperaments, and interests." Effective conceptions of the good must also "suit us as autonomous agents—creatures who engage in self-reflection, who care what we are like and how we are motivated, who seek to define ourselves by what we choose."[43]

[40] Rosati 2007, 232–33
[41] 2007, 234
[42] 2007, 234–35
[43] 2007, 251–52

Rosati distinguishes three kinds of regret, two of which are well-grounded by reference, in different ways, to one's effective conception of the good.

The first and most straightforward kind of well-grounded regret is a response to what we might call *mistakes*—choices that fail to satisfy one's effective conception of the good construed as "correctness conditions" for choosing. Sometimes, however, conceptions of the good are not effective because, for instance, they involve "aims that are, as a practical matter incompatible, or aims that will continually lead to disappointment."[44] A second kind of regret, which we might call *anomic regret*, can be well-grounded as a response to this situation. Finally, Rosati writes of well-grounded "choice regret," which concerns "having to choose and to lose an option we regard as desirable, and such regret involves no sense that we have chosen mistakenly all things considered."[45] The thought here seems to be that choice regret is well-grounded because it is the way the regret subject appreciates the opportunity cost of her choices, including the opportunity cost of choices that are correct according to her effective conception of the good.

I am not certain that choice regret is really regret rather than some other attitude, as it is not obvious to me that the necessity of choosing between desirable ends, or of lacking a basis for correct choice, are possible contents of a regret attitude. Granted, a lifeguard who saves as many people as she could from a large drowning group might naturally say, "I regret not saving everyone." But 'regret' in this case seems to refer to a distinct attitude—a wish, perhaps, that saving everyone had been possible. For it is intuitive to conclude that people do not really regret actions that they would repeat, and endorse repeating, in relevantly similar contexts. Assuming

[44] 2007, 254
[45] 2007, 247

that the lifeguard would again seek to save only as many people as she safely could, even short of everyone, then in a sense—the sense I am interested in discussing—she does *not* regret not saving everyone. Now Rosati may simply be discussing a broader concept of regret than I am, in which case I only observe that I wish to confine my attention to the narrower idea.

Perhaps anomic regret is in fact regret, but if it is distinct from mistake regret, I am not sure how. Presumably, it is not merely mistake regret given a conception of the good that is ineffective on account of producing jointly unsatisfiable correctness conditions for choice. While such a rationally incoherent conception of the good would have the tragic consequence of *inevitable* mistake regret, conceptions of the good can be ineffective in other ways, too: they can fail to suit an agent's current "abilities, temperaments, and interests," or fail to be chosen or endorsed by the agent, and thus fail to suit the agent's autonomy. Conceivably, we might impute to all agents a higher-order desire that their desires be organized in a way that suits them (or suits them in a rationally coherent way)—a higher-order desire, in other words, to have an effective lower-order conception of the good. Consequently, all possible effective conceptions of the good (where we now speak generally of conceptions comprehending desires of every order) include this desire. Well-grounded anomic regret then turns out to be mistake regret well-grounded on *any* conception of the good effective for a given agent. The obvious alternative, of construing anomic regret as basically distinct from mistake regret, returns us to the question of how it can be well-grounded—of what distinguishes anomic regret that is well-grounded from anomic regret that isn't. It would not be enough to simply reply that anomic regret is well-grounded when it is occasioned by the absence of an effective conception of the good, since it seems perfectly reasonable to ask whether one should regret not knowing what one wants.

Most compelling is Rosati's account of the well-groundedness of mistake regret. Mistake regret seems to be regret in the way choice regret does not, as it makes sense to think of mistakes, rather than inevitable opportunity costs, as things that could have been otherwise. In addition, we can clearly determine when it is well-grounded, namely, when it concerns a choice that undermined one's effective conception of the good. This is unlike anomic regret, whose well-groundedness conditions are mysterious unless anomic regret is really a species or extension of mistake regret.

My own attempt to rationalize agent regret is similar in several ways to Rosati's account of mistake regret's well-groundedness. I will presently rationalize agent regret using a principle of instrumental rationality. Like an effective conception of the good, instrumental rationality rationalizes choices—and is therefore grounds for criticizing them—in agent-relative terms. While the norm of instrumental rationality I appeal to is minimal, it can be enriched by other structural constraints on practical reasoning, also in the spirit of rational coherence, in order to achieve the organization of ends constitutive of effective conceptions of the good.[46] As noted earlier, Rosati seeks only to explain how the judgments that regret *involves* can be well-grounded. She distinguishes from this task the question of regret's "appropriateness," that is, the question of whether it is rational to experience regret's affective component.[47] I would like to go beyond Rosati to offer a unified account that rationalizes both components.

[46] I have in mind instrumental principles for weighting or time-ordering of ends, as mentioned by Williams in "Internal and External Reasons" (104). Speaking of internal reasons, instrumental principles combined with reasons internalism, applied retrospectively to evaluate choices, yields a justificatory account of regret similar to Rosati's account involving the retrospective evaluation of choices against an effective conception of the good suited to the agent.

[47] Rosati 2007, 232

Part 2: The Practical Reason Account of Agent Regret

VII. Instrumental Rationality Preliminaries

The Volitional Coherence Principle

An agent can be instrumentally coherent or incoherent with respect to her volitions. For instance, she is incoherent in this way if she wills two mutually exclusive ends. Among the many considerations doubtless bearing on an agent's volitional coherence, I call attention to one that concerns the relationship between means and ends. I will call this, simply, the *volitional coherence principle,* since it is the only one that matters here:

> *Volitional coherence principle:* it ought to be the case for an agent A that: if A wills that φ and A knows that ψ is the means to φ, then A wills that ψ.

Typically, there is more than one known means to a given end, and different means often come with different advantages and different implications for an agent's other ends. These complications are certainly relevant to instrumental reasoning, but for the purposes of this discussion, let us suppose that A has ruled out all other known means ψ' using some sufficiently rationally procedure: perhaps ψ is optimal, or optimal during the period of A's deliberation, or the first satisfactory means occurring to A, or something else of that nature. What is relevant here is that A has already settled on a means ψ to φ—in that sense *the* means—and that, intuitively, there is something irrational about not willing the means when one wills the end. For instance, if it is really my will to apply for a job, and the means to this end is to drop the completed application, now in my hand, into the mailbox several feet away, then I would be rationally incoherent if I did not also have the will to act accordingly. The volitional coherence principle has *wide scope*, which is to say that the *ought* has in its scope the entire conditional (hence the

awkward phrasing), rather than the consequent alone. This means that one can conform to the principle in two different ways: by willing the means to one's willed end, and also by giving up the end. To resume the example, I would cease to run afoul of this principle either by mailing the application or by giving up on applying for the job. This reflects the fact that incoherence between two things is something that can be resolved by adjusting either. I will be assuming this principle as necessary condition of an agent's instrumental rationality. [48]

Instrumental and Final Volitions

It will be convenient to have a term designating a *willing* of some means ψ to a given willed end φ, as well as a term designating the *willing* of the end φ. Let us call the willing of ψ an *instrumental volition* and the willing of φ a *final volition*. In these terms, for example, the volitional coherence principle states that if an agent has a final volition, then he ought to have the

[48] Broome 1999; Schroeder 2004. It is controversial whether and how principles of instrumental rationality are normative. Typically, intentional coherence principles are discussed, such as that an agent who intends an end ought to intend what he knows or believes to be the means. A "narrow-scope" version of this principle, of the form $P \rightarrow O(Q)$, where the *ought,* O, has only the consequent of the conditional in its scope, is immediately problematic. For instance, P may be that an agent intends to commit murder, while Q is the known necessary means, namely, stabbing the victim in the heart. Yet it is absurd that the agent ought to stab his victim in the heart just because that he decides to commit murder (see, e.g., Raz 2005, 2-3; Schroeder 2009, 224). Surely an *ought* cannot be bootstrapped in this way, much less this *ought*.

A wide scope intentional coherence principle of the form $O(P \rightarrow Q)$ avoids this problem, as it blocks detachment of the consequent by modus ponens (Broome 1999, 401-403). Instead, it only follows that the agent ought *either* to intend to stab *or* to cease to intend to murder: $O(\sim P \vee Q)$, which is consistent with giving up the intention to murder being, in fact, the only thing the agent ought to do. Against this, Raz and Schroeder have each argued, on different grounds, that wide scoping does not actually avoid the bootstrapping problem. Raz's argument seems to depend on two questionable entailments. Where R is a *pro tanto* reason operator: $R(P \rightarrow Q) \vdash R(\sim P) \wedge R(Q)$, and also $O(P \rightarrow Q) \vdash R(\sim P) \wedge R(Q)$ (Raz 2005, 12). Yet although Raz's criticism is aimed at Broome, nothing Broome says requires these entailments. Schroeder's criticism, in turn, seems to depend on treating epistemic and practical *oughts* as logically equivalent in the otherwise unimpeachable inference $O(P \rightarrow Q), O(P) \vdash O(Q)$ (Schroeder 2009, 227). This seems implausible. It is not obvious that epistemic and practical *oughts* even range over all of the same objects. And even if they do, or even if we restrict our attention to the intersection of their ranges, it is not obvious that *oughts* from different normative domains would even be consistent with each other.

associated instrumental volition. To return to our example above, the final volition is that I eat dinner at the popular restaurant sometime next week, and the instrumental volition is that I make a reservation.

VIII. Temporal Features of Volitions

In this section I explore two temporal features of volitions, namely, *duration* and *propositional tense*. A mismatch between duration and propositional tense can produce the phenomenon of *disappointment*, which is central to my account of agent regret.

Duration vs. Propositional Tense

Volitions have *duration.* By this I mean that volitions *qua* mental states begin—they are formed—and they conclude—they unconsciously dissipate or are consciously revoked. For instance, if I will that I go to the bank and then later either change my mind or entirely and permanently forget this volition, then the volition's duration is the span of time from forming it to revoking or losing it.

Second, volitions have *propositional tense.* Like all intentional attitudes, volitions have propositional content. This content, in turn, often has a temporal truth condition. For example, if an agent wills that he have coffee during his break from 2-2:15, then the volition's propositional content is only true if he has coffee during that fifteen-minute span. While the volition itself is not true or false, the temporal truth condition of its propositional content—the volition's propositional tense—is nonetheless important, as it is the time when the willed event is willed to take place. If he gets coffee after leaving work at 5 pm, he does not do what he originally willed.

There are many possible propositional tenses. One such tense might span the time between two moments, such as making coffee any time within the next fifteen minutes. Another propositional tense might include but a single moment, such as igniting the rocket engines in exactly ten seconds. A propositional tense might have an indefinite beginning or end—roughly corresponding to grammatical aspect—such as willing that one writes that letter to Aunt Susan sometime in the near future, or willing that one pleases one's parents for as long as one can remember.[49]

Propositional Tense and Disappointment

I call the degree to which a volition's content fails to obtain during its propositional tense the degree to which the volition is *disappointed.* For instance, if an agent wills that he make coffee within fifteen minutes and fails to make it in time, then his volition is totally disappointed. But if instead he wills, say, that he drink no more than two cups of coffee a day for the next month, and on six occasions in the next month he drinks three cups of coffee, then we may say that his volition is partially disappointed (perhaps pending some finer-grained report of the propositional tense of the volition). I shall use 'disappointed' to describe both total and partial disappointment.

Pre-Disappointment

Some volitions have propositional tenses beginning or even ending prior to their duration: I might will that I *had not slept* through my alarm, or that I *had not struck* a child while driving a truck. When reporting such past-oriented wills, we tend to use locutions involving *would*: "If only I'd known the market was going to crash I would have sold all my stock," or, "If I could do it all

[49] In other contexts we might equivalently say "from time immemorial."

over I wouldn't become a lawyer," or, "If I could go back in time, I would take voice lessons instead of slaving away at the violin." A 19th century heroine might say, "Would that I'd never married that heartless rake."

In such volitions, the beginnings of their propositional tenses precede the beginnings of their durations. Consequently, they are typically (though not invariably) *already* partially or totally disappointed.[50] I wish to consider two such cases of *pre-disappointment*. The first is where one *could have formed* the instrumental volition prior to disappointment, and the second is where one *couldn't have formed* it prior to disappointment. [51] As an example of the first case, let us suppose that an agent is firmly committed to not hurting people. Not racing other drivers in residential neighborhoods is a means to this end, but a means the agent fails to will in a moment of weakness. He then strikes and seriously injures someone in a crosswalk. I postpone discussion of the second case until the end of this section as it is similar in most respects to this first case.

Let us suppose that the agent is a *moral* agent. I take the following to be uncontroversial: (1) if he drives in the future, he doesn't race in residential neighborhoods on account of not wanting to hurt people, and (2) he would assent to the statement, "If I could go back in time I would never have raced in residential neighborhoods." Both (1) and (2) can be explained by a single volition: the volition not to race in residential neighborhoods—a volition whose propositional tense begins prior to injuring the pedestrian and extends through the present and

[50] Conceivably, I might *now* form the volition to ψ at some point in the span of three days beginning with yesterday and ending tomorrow, as perhaps when I today discover that yesterday, today, and tomorrow are the days during which I am permitted submit a form.

[51] I am using 'could' as follows: an agent A could have formed a volition to φ at time t if and only if A had reason to φ at t given A's ends at t and A's knowledge at t. I realize that this is not a straightforward epistemic modality. I mean to capture the sense of 'couldn't' (and thus of 'could') in "I couldn't have known!" uttered in self-exculpation.

indefinitely into the future, but whose duration begins only after injuring the pedestrian. It seems that after the accident he forms the very instrumental volition he ought to have had ever since he started driving. From his perspective as a new driver, he ought to will that he not race in residential neighborhoods over the course of his driving career. From his perspective after the accident, he ought to will the very same thing. From both perspectives the volitions are the same, even though from the second perspective part of his driving career is now in the past. The difference in perspective does not change the volition's content—specifically its propositional tense—in the slightest.

We can now consider the second case of pre-disappointment. In this new case the agent *couldn't have formed* the instrumental volition prior to disappointment, whereas we recall that in the first case the agent *could have*. For instance, suppose that the agent hits someone not because he drives at reckless speeds through a residential neighborhood—in fact he is driving very carefully—but because a child suddenly darts out in front of his truck. In this case, the agent couldn't have known that the child would be there. Nonetheless, he wills after the fact that he had made some choice that would not have led him into the child's path: he wills, for instance, that he had driven especially slowly down that road. The difference between this second case and the first is as follows. In the first case, the agent *always knew* that not racing through residential neighborhoods is a means to the end of not hurting people. In the second case, the agent *did not and could not have known* that driving especially slowly on the child's road is a means to the end of not hurting people.

Disappointment Does Not Distinguish Guilt from Innocence

There is a timeless causal relation between not racing through residential neighborhoods and not hurting people, and likewise between driving especially slowly on the child's road and not hurting people. That is to say, one might know that these relations obtain *before* they bear on one's decisions—*before* one races in a residential neighborhood (or drives normally on the child's road)—and one might know them *after* they bear on one's decisions—*after* one strikes the pedestrian (or the child). The fact of these relations obtaining does not change over time. Consequently, by the volitional coherence principle, an agent who wills that he not hurt people *ought* to have the relevant instrumental volition *as soon as he knows that the causal relation between means and ends obtains*, even if it is too late to carry that volition out—even if the volition is pre-disappointed.

It might seem that the volitional coherence principle cannot require pre-disappointed instrumental volitions, since such volitions cannot be carried out in service of the agent's final volitions. The volitional coherence principle does in fact require them, however. As we recall, the principle states that if an agent A wills that φ, if ψ-ing is the means to φ, and if A knows that ψ-ing is the means to φ, then A ought to will to ψ. It is true that the volition to ψ—the mental state—is no longer a means to φ. But that is irrelevant: by the volitional coherence principle, one ought to will to ψ in order to φ not because the *volition* to ψ is a means to φ, but because ψ-ing is a means to φ. The causal relation between ψ-ing in a given span of time and φ obtaining in a

given span of time is itself a timeless metaphysical fact, one independent of the time at which an agent forms volitions concerning φ or ψ. [52]

In conclusion, even if one is innocent, one can be rationally obliged to have a past-oriented volition *just as if* one were guilty: both the innocent and guilty drivers might will that they had not driven as they drove—the innocent driver wills that he had driven especially slowly, and the guilty driver wills that he had not driven so fast. The innocent driver now has a past-tense volition just like the guilty driver, and the innocent driver's innocence is not entailed or even necessarily registered in any way by his volition's propositional tense or content. The fact of disappointment is the same for both agents. The only difference between them is that only the guilty agent has actually *violated* the volitional coherence principle: only the guilty agent *knew in advance* of the instrumental relation between not doing what he did and not hurting people, and so only the guilty agent was required to have had the corresponding instrumental volition *in advance* of his accident.

[52] I implicitly granted here that past-oriented pre-disappointed volitions *qua* mental states may well be instrumentally futile, but this need not be the case: they can play an important role in an agent's future-oriented projects. For when changes occur in the agent's knowledge of what means serve what ends, the set of instrumentally required volitions changes as well. Even if the newly acquired volitions are pre-disappointed, they can usefully bear on the agent's future plans. For instance, suppose a surgeon operating on a patient's heart experiences a hand tremor, which she recognizes as a symptom of incipient Parkinson's disease. Her scalpel slips from her fingers and into the patient, killing him. She now knows that not operating on people is instrumental to her end of doing no harm, and because she is rational, she wills that she not have operated. This volition has a propositional tense beginning in the past but extending into the future. The volition is partially disappointed because she already operated, but since its propositional tense extends beyond the operation and into the future, it moves her to give up surgery. In this manner, self-reproach for something past causes her to alter her future plans in a way instrumental to her ends.

IX. The Appropriateness of Agent Regret

The goal of this paper, we recall, is to answer the following question: is agent regret, understood as moral self-reproach, an appropriate response to causing harm through no fault of one's own? In this section I lay out my answer, which I illustrate by returning to Williams's truck driver.

Let us assume that the driver is a moral agent, and therefore has the end of not killing children. That is to say, not killing children is an end he adopts for moral reasons. By the volitional coherence principle, he is *rationally required* to will that he drive especially slowly on the child's street, since he knows *now* that by doing so he will disappoint his end of not killing children. Since this volition is rationally required, it is appropriate. Since the fateful day is past, the volition is pre-disappointed. Since the volition is pre-disappointed and since he is a rational agent, I propose that he experiences his volition as self-reproach, as I will now explain.

Consider that the driver comes to know of the instrumental relationship between driving especially slowly on that street *and* not killing children *only by driving normally down that street and killing a child.* He thus finds himself vividly confronted by irrefutable evidence of the instrumental necessity of willing that he drive especially slowly there, and because he is rational he so wills. But this will is already frustrated: he finds himself futilely striving to change his involvement in events now forever beyond his control. This tension, which is brought on by the conative character of volition and the unachievable character of the volition's object, together account for the self-reproachful character of his experience. Finally, the self-reproachful character of this volition is *moral* because the volition is instrumental to a moral end.

By accounting for the self-reproachful character of agent regret, I avoid the objection I raised against Tannenbaum. We recall that she does not explain how blamelessly causing harm justifies *agent regret* rather than some other, phenomenally distinct feeling. She cannot rule out even feelings that are entirely inappropriate, such as glee, because she understands agent regret as a *feeling associated with a judgment*, and it seems possible to associate a variety of feelings with a judgment. It is much less plausible, however, that glee would be the experience of the tension between the conative character of volition and the unachievable character of the volition's object, as I would have it.

In addition, by identifying agent regret as the experience of this tension, I avoid the objection I raised against Helmreich. We recall that he doesn't establish that it is appropriate for all moral agents to experience agent regret for blamelessly causing harm. He does show that agent regret is appropriate insofar as the agent's odds of success are increased by adopting a self-critical orientation towards failure. He does not, however, then establish that such an orientation contributes to success for all moral agents, and so he does not establish that agent regret is for all moral agents appropriate. In contrast, my account succeeds in showing this because all moral agents who blamelessly cause harm are *rationally required* to have volitions they now cannot carry out, and having those volitions is experienced as agent regret. And if all moral agents who blamelessly cause harm are rationally required to have the response of agent regret, then *a fortiori*, it is *appropriate* for all such agents to experience agent regret.

Furthermore, the timeless character of the instrumental relation between means and ends allows my account to be more general than Sussman's. As we recall, Sussman rationalizes agent regret by appealing to the social function of making amends—to how the truck driver's quasi-

apology, together with the pedestrian's quasi-forgiveness, allow them to return to the social status quo ante. But where no amends can be made, as when the pedestrian is killed rather than injured and where there are no friends or relations to accept the quasi-apology instead, it seems that making amends is impossible. Sussman's account does not appear to rationalize agent regret in such cases, which my account is able to cover by rationalizing agent regret solely in terms of instrumental rationality and the agent's own ends.

Like Raz, I maintain that agent regret is fundamentally self-reflexive: the subject experiencing agent regret is the same person as the agent responsible for the action occasioning agent regret. As discussed earlier, Wallace rebuts Raz's claim that agent regret is distinct from spectator regret because agent regret's self-reflexivity gives it unique content. While Wallace may be right that spectator regret can also have this content, the possibility that agent regret is just a case of spectator regret is undermined by his own analogizing of agent regret with intention. By unreservedly incorporating the insight of this analogy, my account is able to support the distinctness of agent regret while also explaining the orectic attitude of the regret subject—that is, the conative character of the regret attitude—that Wallace usefully identifies.

Finally, my account incorporates Rosati's thoughts about regret's justificatory structure. My rationalization of agent regret depends on instrumental rationality in a way conscious of how her account of the well-groundedness of mistake regret depends on effective conceptions of the good. My account learns from her discussion in order to go beyond its stated scope and to rationalize the affective component of regret in addition to the cognitive component that Rosati insightfully discusses.

X. Conclusion

I hope to have shown that it is appropriate for moral agents who blamelessly caused harm to experience agent regret understood as moral self-reproach. I began by discussing Frankfurt and Sussman, who deny that agent regret thus understood is appropriate at all. I next considered Tannenbaum, Helmreich, Raz, Wallace, and Rosati, who agree that agent regret thus understood is appropriate, but on grounds I found problematic. I then laid the foundation for my own positive account claim by assuming a minimal conception of practical reason as given by what I called the volitional coherence principle. Next I observed that volitions have the temporal features of propositional tense and duration. It was significant when the beginning of a volition's propositional tense precedes the beginning of its duration, since such volitions might be impossible to execute. It turned out that the volitional coherence principle requires such volitions of a moral agent who blamelessly causes harm. I then argued that agents experience as self-reproach the tension between having a volition and knowing that it cannot be carried out. When the volition in question is a means to an end adopted for moral reasons, such as not causing harm, the agent's self-reproach is *moral* self-reproach. It followed that moral agents who blamelessly caused harm are rationally required to have volitions they cannot but experience in a morally self-reproaching way. Moral self-reproach is thus not only appropriate but rationally required when moral agent cause harm even when they are morally blameless for causing it.

Chapter 2: Making Them Regret It

Volitional Empathy as a Success Condition of Punishment

A man is at a bar on a Friday night after a long week. He encounters another patron and their conversation quickly sours. Words are exchanged. Blows seem sure to follow, but our man storms out of the establishment. Later that night, outside the bar and around the corner, he ambushes his antagonist with a pipe wrench and bludgeons him to death. It doesn't take long for the police to apprehend him: his victim, after all, was found not far from the bar where many other patrons witnessed the altercation that night. He is charged with murder. He is given a procedurally fair trial, he is convicted, and he is sentenced appropriately—perhaps to many years in prison, or perhaps to be "hanged from the neck until dead."

When he is led from the courtroom, he curses the judge, the jury, and the lawyers. He stews in his vindictive anger until the day appointed for his execution. On the gallows he curses the executioner. He smiles maliciously at the victim's family at the front of the onlooking crowd. He says that he will see them all in Hell. Or perhaps, having been sentenced to prison instead of death, he whiles away his long sentence dreaming of revenge. He views his incarceration as a kind of amoral outrage against his person—a violation that superior power has compelled him to suffer, and one to be avenged by violations that he dearly hopes will one day be his turn to inflict. Perhaps he never sees this through when he is finally released, but only out of an abundance of prudence. Let us call someone who responds to punishment in this way a *vengeful villain*.

Or instead, perhaps the assailant is dragged from the courtroom shouting with sincere righteous indignation that he goes now to be butchered on the altar of tyranny. For you see, his victim was a soldier, and soldiers, being agents of state violence, are in his view the criminal instruments of a despotic regime. The action was not murder—or at least not in any morally problematic sense of the word—but rather a blow struck for justice and revolution. He sees himself as a political prisoner, and he is convinced that he will be vindicated in the judgment of history. Or similarly, perhaps he sees the government as an irreligious abomination. He goes to his punishment with not only a clear conscience but conspicuous pride, innocent as he is in the eyes of God. Let us call someone who responds to punishment in these ways a *martyr*.

Or instead, perhaps as he goes from the courtroom, there is a distinctive spring in his step—almost a lurch—betraying the secret delight with which he beholds many years in what he sees as a taxpayer-funded paradise of masochistic delights. And in fact, his sentence, which would in most others cause an acute sense of deprivation, in his case interacts with his predilections to cause the utmost fulfillment. Or perhaps he goes to his execution eagerly expecting ecstasies beyond description in the moments before he expires. Let us call someone who responds to punishment in this way a *masochist*.

I suspect many would agree that there is something wrong with a punishment when the person to whom it is administered responds in one of these ways. This is not to say that such a punishment is not, in fact, punishment, but that it is somehow *incomplete* or *deficient*.[53] I propose in this paper to explain this deficiency as the failure of punishment to induce the wrongdoer to

[53] I set aside the question of whether deficient punishment is still punishment. My discussion is consistent with deficient punishment being punishment, but failing to accomplish one of its purposes; and also with deficient punishment being punishment-like, but not strictly punishment.

experience a species of moral regret involving what I call *volitional empathy.* That is to say, punishment fails in our problem cases because it fails to provoke the wrongdoer's empathetic participation in the volitional states of his victim.[54] In the first section of this paper, I will show that traditional desiderata for successful punishment struggle to explain how punishment in our problem cases is deficient. In the second section, I develop the concept of moral regret arising from volitional empathy—a species of regret that I call *empathetic moral regret*—and whose absence, I propose, explains the deficiency of punishment in our problem cases.[55] Finally, I will discuss how empathetic moral regret may be not merely a necessary condition for punishment's success but the very purpose of the experience of punishment.

I. The insufficiency of other desiderata

In this section I consider whether the deficiency in the punishment of our three problem cases can be explained by the failure to achieve one of three desiderata put forward by existing theories of punishment. These desiderata include giving wrongdoers their moral deserts as proposed by retributivist theories, the moral education of the wrongdoer and the public as proposed by Jean Hampton and Herbert Morris, and the expression of condemnation as proposed by Joel Feinberg.[56]

---

[54] My discussion does not apply to punishment of wrongs where no one is harmed in any sense, directly or indirectly. Perhaps someone with no family or friends who harms himself in a way prohibited by law would be such a "wrongdoer," or at least a criminal.

[55] This proposal only attempts to identify something missing from the theories of punishment to be discussed in the paper's first section, namely, how punishment should affect the wrongdoer. I am not offering an alternative theory of punishment, which would also need to explain various other features of punishment—for instance, the relation between punishment and law.

[56] I set aside the desideratum of deterrence because I wish to examine the subjective response of wrongdoers to punishment (as distinct from behavioral reform of the wrongdoer or of spectators), and because I take the problem cases to show, if anything, that a wrongdoer's subjective response can affect punishment's success.

A. The retributive desideratum

I will first consider whether achieving the aim of retributive theories of punishment can rectify the unsuccessful cases of punishment initially described. Before I begin, however, I set aside the important question of whether retributive theories are morally supportable. I also set aside this question when we discuss other punishment desiderata, but I wish to make special note of this matter here because the question of moral justification is especially urgent when a theory demands that we punish by retaliating[57] harms against wrongdoers. I table this question because the issue under consideration is whether punishment is successful, and not whether it is morally justifiable.

According to H.L.A Hart, retributive theories of punishment make three claims:

> [F]irst, that a person may be punished if, and only if, he has voluntarily done something morally wrong; secondly, that his punishment must in some way match, or be the equivalent of, the wickedness of his offence; and thirdly, that the justification for punishing men under such conditions is that the return of suffering for moral evil voluntarily done, is itself just or morally good.[58]

Hugo Bedau has observed that this model of retributive theories omits what we might imagine to be the central desideratum of retributive punishment: giving wrongdoers what they *deserve*. So essential is the concept of desert, according to Bedau, that "retributivism without *desert*—the concept of punishment as something deserved by whoever is rightly liable to it—is like *Hamlet* without the Prince of Denmark".[59] Bedau offers several conservative suggestions for

[57] 'Retaliation' in the sense of *lex talionis*—as returning harms somehow *in kind*—and not in the sense of simple revenge.
[58] Hart 2008, 231
[59] Bedau 1978, 608

how the concept of desert might be supplied, such as by making Hart's first and second claims the truth conditions for propositions involving desert, or by requiring that punishment be "imposed by reference to" a criminal's wrongdoing, in which case it is deserved if it is properly proportional.[60] In any event, Bedau does not seem especially worried by this omission of Hart's, and the fact of the omission testifies to Hart's own lack of concern. I will take Bedau's suggestion that desert's truth conditions are given by Hart's first and second claims. This is convenient for our purposes since we are tabling the issue of moral justification, and that issue is what Hart's third claim concerns. Desert, then, is whatever is supplied by a punishment that appropriately matches the "wickedness" of a wrong voluntarily committed by the person punished. Let this be the retributive desideratum for us to examine.

Next to time-honored if unfashionably rigorous sentences involving eyes and teeth, the most famous examples of retributive punishment are probably those described by Kant in the *Metaphysics of Morals.* Kant's first two examples seem aimed to show how injuries can be retaliated in kind, with appropriate adjustments, even when the circumstances of the perpetrator and victim crucially differ. But no such complications arise in the case of murder, for which the only appropriate punishment is apparently death. Kant takes the requital of death for murder so seriously that he famously insists that it is the duty of the state to carry out all death sentences even if it were about to dissolve itself, so that "each has done to him what his deeds deserve and blood guilt to does not cling to the people for not having insisted upon this punishment."[61]

[60] Bedau 1978, 609
[61] Kant 2017, 6:332-3

Since the vengeful villain, that masochist, and the martyr are all guilty of murder, death, then, is the punishment whose success we should evaluate, according to Kant. Let us suppose that Kant is right, and that death is indeed what our wrongdoers deserve. It is certain that many have gone to their executions with the vindictive attitude of the vengeful villain and the self-righteous attitude martyr, and there is little reason to think that the strange attitude of the masochist is outside the realm of psychological possibility. Is, then, the punishment of death successful when the perpetrators to be punished bear towards their punishment these problematic attitudes?

Let us consider the attitudes of our three perpetrators more closely. The vengeful villain believes that his punishment is an evil that he does not deserve to suffer, and even if he comes to accept his crime as a wrong, he objects to his punishment as nothing more than revenge. The masochist, in contrast, believes that deserved or not, his punishment will not only not involve suffering, but will be a delight, and he looks forward to it with pleasure rather than dread. Finally, the martyr, like the vengeful villain, believes that his punishment is an evil that he does not deserve to suffer, and like the masochist, he looks forward to his punishment. He does not, however, look forward to his punishment as a pleasure, but as a moral victory over the punishing authority. In short, the vengeful villain has inappropriate beliefs about desert, the masochist has inappropriate beliefs about suffering, and the martyr has inappropriate beliefs about the kind of moral good achieved by punishing him.

I grant that one may consistently hold that the central claims of retributivism are satisfied by such punishment *and* that the punishment is successful. However, at this point I hope to have clarified the intuition that in the punishment of these perpetrators, something important is missing. This intuition reflects an argumentative burden that anyone must meet who holds that

satisfying the central claims of retributivism is sufficient for the success of punishment. That burden is to explain how punishment can succeed irrespective of how it is received by the punished.

B. The expressive desideratum

The next desideratum I would like to consider is what Feinberg has called "the expressive function of punishment" in his article of the same name. Feinberg describes this function as follows:

> Punishment is a conventional device for the expression of attitudes of resentment and indignation, and of judgments of disapproval and reprobation, either on the part of the punishing authority himself or of those "in whose name" the punishment is inflicted.[62]

Feinberg later clarifies that reprobation just *is* the "stern judgment of disapproval," and he seems to subsume indignation, along with other and "various vengeful attitudes," under the name of 'resentment'. We may therefore distinguish two features of the expressive function:

(1) *Resentment*: various vengeful attitudes

(2) *Reprobation*: a stern judgment of disapproval.

The expression of both attitudes, which Feinberg calls *condemnation,* constitutes the *primary* expressive function of punishment.[63] I say 'primary' because Feinberg proceeds to list four "derivative symbolic" functions of punishment that presuppose condemnation. They are as follows:

[62] Feinberg 1965, 400
[63] Feinberg 1965, 403

(3) *Authoritative disavowal*: the execution of punishment supports the punishing authority's additional claim that the punished party was not acting on the authority's behalf.

(4) S*ymbolic non-acquiescence*: the execution of punishment discharges the responsibility of the people authorizing the punishing authority (such as citizens where the authority is their government) to recognize a wrong as such.

(5) *Vindication of the law*: the execution of punishment defends the mandatory character essential to the law—a character that erodes when laws are not enforced by punishments. As Feinberg puts it, "a statute honored mainly in the breach begins to lose its character as law, unless, as we say, it is *vindicated* (emphatically reaffirmed); and clearly the way to do this (indeed the only way) is to punish those who violate it."

(6) *Absolution of others*: where relevant, the execution of punishment exonerates other suspects of the same wrongdoing.[64]

We need not add anything out of the ordinary to the punishments of the vengeful villain, the masochist, and the martyr to easily satisfy the expressive conditions as given. Indeed, it would be hard to see how a punishment prescribed by law could fail to satisfy these conditions so long as it is sufficiently severe and public, and so long as it is pronounced and executed with official pomp. That said, we might make the satisfaction of these conditions vivid by imagining the robed and bewigged president of a tribunal, enthroned on a high platform before the seal of the state and the ensigns of the law, gravely pronouncing the sentence of death upon the offender. The condemned man is conducted to a public square where there awaits a firing squad comprised of ten citizens randomly selected for the duty as they might have been for a jury. As the man is

[64] Feinberg 1965, 404–8

hooded, his crimes are declaimed to a large and expectant crowd, and his death is solemnly devoted to Justice and to Peace. The order is given, the shots ring out, the man collapses dead into his chains. After a moment of silence, the crowd erupts in cathartic applause.

Before we can consider whether the expressive desideratum can explain the deficiency of the punishment in our problematic cases, we need to distinguish the expressive and retributive desiderata. We may thereby consider the expressive desideratum's unique explanatory contribution, which might in turn conceivably explain the deficiency in punishment in our problem cases. This clarification is needed because retributive and expressive desiderata substantially overlap. We can illuminate the boundary between these desiderata by considering the vengeful villain, for whom both retributive and expressive punishment seems especially apt—even more so than for the masochist or the martyr.

Ex hypothesi, the vengeful villain, the masochist, and the martyr have committed identical wrongs—they differ only in their attitudes towards their sentences. With respect to punishment's reprobative function and the other derivative functions, it is far from clear how punishment of the vengeful villain is especially fitting. Since there is no difference in the wrong or its circumstances committed by the three wrongdoers, there are no grounds for the "stern judgment of disapproval" being different in content or sternness for the vengeful villain. The same can be said for the derivative functions symbolic non-acquiescence and absolution of others. And if anything, the vengeful villain's insolent attitude makes authoritative disavowal *less* necessary, since it is less necessary for the state to support with punishment its claim that the vengeful villain was not acting on its behalf when it can also cite his contempt for its agents. Perhaps there is greater need to "vindicate" the law in the face of contempt. But insofar as it is the failure of the relevant

authority to *enforce* of the law and not the contempt of the lawbreaker that erodes a law's character as such—as Feinberg suggests when he gives the example of racist jury nullification of murder laws during the prosecution of white perpetrators with black victims—if indeed it is failure of enforcement that erodes a law's lawlike character, then the vengeful villain's insolent attitude is irrelevant.[65]

The expression of resentment is the only remaining function to which we might appeal to explain the superior aptness of punishing the vengeful villain. Feinberg does not develop this feature of punishment in much detail. In fact, he is clearly reluctant to include resentment among punishment's expressive functions, and does so only with a frown and a sigh:

> At its best, in civilized and democratic countries, punishment surely expresses the community's strong *disapproval* of what the criminal did. Indeed, it can be said that punishment expresses the *judgment* (as distinct from any emotion) of the community that what the criminal did was wrong. I think it is fair to say of our community, however, that punishment generally expresses more than judgments of disapproval; it is also a symbolic way of getting back at the criminal, of expressing a kind of vindictive resentment.[66]

Feinberg proceeds, as if with mounting bewilderment, to describe as a vehicle for this resentment the "self-righteous" and "naked hostility" of a prisoner's "guards and the outside world"— treatment which "bears the aspect of legitimized vengefulness."

We seem to have an explanation for why the vengeful villain seems especially punishable: his contempt is insulting, there is nothing more natural than revenging an insult, and one of the

[65] Feinberg 1965, 407
[66] Feinberg 1965, 403

functions of punishment is to express the vengeful attitudes that constitute resentment. But before we can consider whether the expression of resentment, given the other expressive criteria, can make punishment successful, we must take note of two factors that complicate the role expressing resentment plays in punishment. The first is that Feinberg does not seem to think of expressing resentment as conceptually necessary for punishment, since "at its best, in civilized and democratic countries," even if not "in our community," punishment expresses only a "judgment (as distinct from any emotion)." The second is that resentment is not merely "symbolized" by the suffering punishment invariably involves. As A. J. Skillen observes:

> Whereas black is arguably neutral in itself and only contextually and conventionally constituted as mourning wear… it is pretty clear that losing money, years of liberty, or parts of one's body is hardly neutral in that way…. Feinberg vastly underrates the natural appropriateness, the non-arbitrariness, of certain forms of hard treatment to be the expression or communication of moralistic and punitive attitudes. Such practices *embody* punitive hostility, they do not merely "symbolize" it.[67]

There is, in short, a close relationship—one much closer than mere conventional symbolism—between resentment and inflicting suffering. I offer that the relationship in question is that resentment, especially *qua* "various vengeful attitudes," is *satisfied* by inflicting suffering on its object. Feinberg himself suggests as much when, in culminating his short discussion of resentment, he quotes J. F. Stephen's observation that "the criminal law stands to the passion of revenge in much the same relation as marriage to the sexual appetite."[68] Stephen probably did not mean that marriage's relation to the sexual appetite is that it prescribes that appetite's

[67] Skillen 1980, 517
[68] Stephen 1863, 99

conventional, symbolic expression—say, in the composition of love poetry—but rather, that marriage prescribes the satisfaction of the sexual appetite by sexual intercourse. In keeping with the analogy, then, the criminal law's relation to the passion of revenge is not that it prescribes the symbolic expression of the passion of revenge, but rather its satisfaction by the suffering inflicted on the criminal.

Feinberg seems to think that suffering, too, is conceptually unnecessary for punishment. After observing the evolution over time of the "symbols of shame and ignominy" used in punishment, and noting the possibility of "stigmatiz[ing] without inflicting any further (pointless) pain to the body, to family, to creative capacity," Feinberg considers an "elaborate public ritual, exploiting the trustiest devices of religion and mystery, music and drama, to express in the most solemn way the community's condemnation of a criminal for his dastardly deed," in order to "preserve the condemnatory function of punishment while dispensing with its usual physical forms—incarceration and corporal mistreatment."[69]

Feinberg is uncertain whether this hypothetical ritual is an "idle fantasy."[70] It probably is given two things: first, that I am right about the relationship between punishment and resentment's vengeful attitudes—that the former satisfies and does not merely express the latter—since it is hard to see how vengeful attitudes can be satisfied without vengeance; and second: that we interpret Feinberg as offering a hybrid account of the function of punishment, namely, to express reprobation and to *satisfy* (and perhaps also to express) resentment. On the other hand, Feinberg's ritual is probably not an idle fantasy if we take Feinberg's language of

[69] Feinberg 1965, 420
[70] Feinberg 1965, 421

expression as meaning just that and nothing more. I would venture to diagnose Feinberg's ambivalence about the possibility of punishment without suffering as a symptom of conflating the expression of resentment with its satisfaction.

If we interpret Feinberg as insisting on resentment's satisfaction, which I take to be a retributive aim, then I have nothing to add here to my discussion in the previous section of the retributive desideratum. If, however, we interpret Feinberg as insisting only on resentment's expression (if only in a society sufficiently "civilized and democratic"), then we should consider the sort of punishment ritual he has in mind. I am not sure what to make of Feinberg's suggestion that the ritual involve "trustiest devices of religion and mystery, music and drama," which seem more likely to produce aesthetic detachment than to express public resentment. (Perhaps Feinberg felt that in a civilized and democratic society, even resentment's mere expression could be dispensed with.) If anything, the ritual should have the *appearance* of satisfying public resentment—it would express the desire for revenge by *representing* its satisfaction. It would, in short, look like the execution described earlier (or like any other punishment sufficiently severe and public to satisfy Feinberg's expressive criteria). We might imagine the condemned criminal—awed or perhaps drugged into compliance—falling on cue as his executioners discharged their blank rounds in his direction. This would not be a mock execution—everyone, including the criminal himself, would know that this ritual is meant to express, without stooping to satisfy, the vengeful attitudes of the punishing authority and the society that authorized it.

Let us suppose that this ritual punishment was carried out upon the vengeful villain, the masochist, and the martyr. Each for entirely nonmoral reasons sincerely resolves never to murder again. Perhaps they found the experience overwhelmingly humiliating, or perhaps they couldn't

bear the disgust they felt at the weakness and hypocrisy of a society that delights in pantomiming brutality it condemns. My point here is not to indulge in speculation about the possible unsatisfactory responses to the punishment ritual described, but rather to point out that the success of punishment intuitively depends to some extent on the response of the punished, and that it is still possible, even when the criteria for expressive punishment are met, for this response to be inadequate.

C. The moral education desideratum

The last desideratum I would like to consider is the moral education of the wrongdoer and of the public. As Hampton puts it, "Punishment is intended as a way of teaching the wrongdoer that the action she did (or wants to do) is forbidden because it is morally wrong and should not be done for that reason. The [moral education] theory also regards that lesson as public, and thus as directed to the rest of society."[71] I take Hampton to mean here that the wrongdoer is taught to forbear from an action because it is wrong, rather than to forbear from an action because it is forbidden.[72] Of the essence of this desideratum is that punishment *teaches moral reasons* to *choose* not to do wrong. The goal of punishment is not, or is not primarily, to *merely* deter antisocial behavior by striking fear or imposing costs.[73] Hampton does allow that a secondary purpose of punishment is to merely deter the "criminal [who] refuses to understand the state's communication about why there is a barrier to his action," though mere deterrence can be distinguished from the primary desideratum of moral education.[74] With respect to the

[71] Hampton 1984, 212
[72] For Hampton, the relationship between the prohibition and the wrong seems to be that the prohibition is a legal representation, made vivid by punishment, of a moral requirement. The audience for this representation consists of those who might not otherwise see the requirement in question (212).
[73] Hampton 1984, 226–27
[74] Hampton 1984, 231

desideratum of moral education, then, we may say that a wrongdoer is successfully punished when two conditions are met:

(i) The punishment agent communicates to the wrongdoer and to the public, by means of the punishment, that the punished action should not be done *because it is wrong*.

(ii) On account of (i), the wrongdoer accepts that the punished action should not be done because it is wrong.

Would the satisfaction of (i) and (ii) transform the deficient punishments of the vengeful villain, the masochist, and the martyr? For several reasons I am doubtful. First, the moral education view diagnoses the moral defect in wrongdoers as a deficiency in knowledge, but only the martyr is clearly mistaken about what is morally required, since the martyr is morally motivated, but in the thrall of the wrong moral theory. The vengeful villain and the masochist *might* be deficient in moral knowledge, but since the motives for their wrongs were nonmoral, it is also possible that their moral knowledge was adequate, but that it does not receive the usual weight during deliberation. That is to say, their problem is that they are selfish because they give excessive weight to nonmoral egocentric considerations, rather than that they are morally solipsistic because they are unaware of the moral considerations to begin with.

The vengeful villain and the masochist, for example, might be disposed to sincerely judge actions of the sort they were punished for as wrong, and to believe that such actions should not be done on account of their wrongness. However, were they to find themselves provoked in the way that they were the night they murdered their victims, they might restrain themselves on account of nonmoral considerations alone, without which they would become repeat offenders. The vengeful villain recalls with impotent rage the jeering of the public during his perp walk. The

masochist recalls the bitter disappointment of compulsory community service, which he completed only to convince his parole board to save him from the endless tedium. Both, in short, were convinced by their punishment that they should not murder because murder is wrong, but neither is ultimately disposed by that judgment to forbear from murdering, at least not in every circumstance where such forbearance is morally required.

The moral education theorist might reply that this objection misconstrues (ii): it is incorrect, perhaps, to consider someone to have moral knowledge who is aware of moral considerations but finds himself unmoved by them. In other words, even if the vengeful villain and the masochist were aware of the relevant moral considerations, the very possibility that such considerations could be outweighed by others betrays moral ignorance—if not of the existence of moral considerations, then of their deliberative role.

I, for my part, am not fully convinced that the weight to be assigned to a consideration is best understood as a kind of knowledge. But let us grant that it is. Let us suppose that the vengeful villain and the masochist both judge that murder is wrong, and that this judgment is sufficient to motivate them to forbear from murdering again. What would they make of their past murders? They would certainly judge those actions as wrong and not to be repeated. But the judgment that one's past action was wrong does not by itself entail that one *regret* one's action. It entails that one should not have acted so. It entails that past deliberation would have properly concluded in the forbearance of murder. It entails, even, that deliberation *must* have properly concluded in the forbearance of murder—that *no possible nonmoral consideration* could have outweighed the moral requirement that murder be avoided. But it seems perfectly consistent for the educated masochist and the vengeful villain to accept all of these judgments and yet recall

their past wrongs with equanimity. After all, they were deficient in moral knowledge, and this predictably led them to commit moral wrongs, which they are now in an epistemic position to avoid.

This brings me to the second reason I am skeptical that satisfying (i) and (ii) would transform the defective punishments of the wrongdoers under consideration. It is intuitively problematic that successful punishment—especially punishment aiming to reform—need not lead the punished to be *bothered* by their past wrongdoing. In circumstances where people suffer some misfortune, especially (though not exclusively) as a result of their own actions, we take their distress as evidence that something they valued was lost, and inversely, we take the absence of distress as evidence that whatever was lost was not of value to them. If someone is distressed over losing a worthless trinket, this is surely because the trinket had some great value to them—that it was not to them nearly so worthless as it seemed to us. Likewise, if someone shrugs at the loss of his life's savings, then it is reasonably to conclude that this person did not subjectively value his life's savings nearly to the extent that we would expect. The intuition that people should be bothered by their wrongdoing is explained, then, by our expectation that people come to value whatever their wrongdoing harmed or destroyed. Equanimity (or surprise or anything other than the right sort of distress) is evidence that this expectation remains unmet.

The moral education theorist might reply, in turn, that whatever it is that wrongdoers should come to value can simply be incorporated into the moral education curriculum. I grant that under a sufficiently expansive concept of education, perhaps this is true. My point is not that the moral education view is deficient under all possible interpretations of 'education', but rather that the concept of education seems to lend itself to teaching people to value things

prospectively—in order to guide their actions—and not *retrospectively,* as grounds for distress when by their actions they harm or destroy valuable things.

Herbert Morris gives a richer description, under the name of 'paternalism', of what Hampton calls 'moral education'. For Morris, the goal of paternalism is "an autonomous individual freely attached to that which is good, those relationships with others that sustain and give meaning to a life." Morris requires that a wrongdoer satisfy a number of overlapping conditions in order to approach this good. Most relevant to our discussion is his first condition:

> First, it is a part of this good that one comes to appreciate the nature of the evil involved for others and for oneself in one's doing wrong. This requires empathy, a putting oneself in another's position; it also requires that imaginative capacity to take in the implications for one's future self of the evil one has done; it further requires an attachment to being a person of a certain kind. The claim is that it is good for the person, and essential to one's status as a moral person, that the evil underlying wrongdoing and the evil radiating from it be comprehended… not merely, if at all, in the sense of one's being able to articulate what one has done, but rather comprehended in the way remorse implies comprehension of the evil caused.[75]

I confess that I would be very hard pressed to describe a punished wrongdoer satisfying this condition—much less Morris's others—whose punishment is intuitively unsuccessful. That said, this condition obscures the relationship between empathy and remorse and lays Morris open to Russ Shafer-Landau's charge that Morris (as well as Hampton) fails to explain how punishment

[75] Morris 1981, 265

is consistent with the claim of Morris and Hampton that their accounts of punishment respect a wrongdoer's autonomy.[76]

Morris thinks that the wrongdoer's understanding of his wrong results from comparing the sort of person who causes in his victim what he empathetically experienced with the sort of person he is committed to being, and finding that the former falls short of the latter. I wholeheartedly agree that successful punishment critically depends on the wrongdoer coming to empathize with his victim. But the *kind* of empathy is of the first importance if it is to do the work that Morris assigns to it—if it is to produce in the wrongdoer the comprehension "not merely, if at all, in the sense of one's being able to articulate what one has done, but rather… in the way remorse implies comprehension of the evil caused." For example, I might value being a good person in the way that a true amateur—a true lover of the sport for its own sake—might value being a good golfer. Since I cannot simply consult a scorecard to determine whether I've wronged someone, I imagine myself in his place, and by this empathetic procedure I determine that my actions have caused another to feel violated. I reflect with disappointment on my moral underperformance, since such behavior is not consistent with the proficient moral athlete I am committed to becoming. But I feel no more remorse for what I've done than the golfer would who is not lately on his game.

In addition, so long as there is no constraint on the sort of person one may autonomously seek to be, then in many (if not most) cases, it will not be possible to punish wrongdoers without violating their autonomy. We need not even consider the person who has no commitment whatsoever to being morally upright. For insofar as people are inclined to provide answers in

[76] Shafer-Landau 1991, 194–97

addition to "morally upright" to the question, "What sort of person would you like to be"—insofar as they are inclined to say "wealthy," "respected," "attractive," "powerful," or "preeminent in bicycle racing"—then compelling people to develop their moral uprightness at the expense of, say, their bicycle racing careers, is a *prima facie* violation of their autonomy.

II. Regret

The punishments of the vengeful villain, the martyr, and the masochist do not fully succeed, as shown by their lack of regret. I will argue that the incompleteness of their punishment is explained by the fact that the wrongdoers are no more responsive, on account of their punishments, to the moral reasons arguing against doing those wrongs than before those wrongs were done. But before I argue this, I wish in this section to clarify the concept of regret that I will be using.

A. Regret is a volitional state

Whatever else it is, regret is an intentional state whose object, in central cases, is the regret subject's own past action or inaction: perhaps I regret not purchasing Facebook stock during its initial public offering; perhaps you regret snapping at a friend while you were in a bad mood. It would be very strange for me to say that I regret your snapping at a friend, or for you to say that you regret my failure to invest in Facebook. I might be *angry* or *sad* or *disappointed* about your action and you might be likewise about my inaction, but *regret* seems quite inappropriate.[77] It would be at least as strange if I now claimed to regret my future failure to invest in the next

[77] Perhaps I might regret your action if you are my agent—I might conceivably regret that, as my attorney, you settled for less than my case was worth—but perhaps this is because your actions *qua* my agent are in a relevant sense my actions too.

successful startup, or for you now to claim to regret your future failure to control your temper that finally and irreparably alienates your long-suffering friend. We might well be *resigned* to a future action or inaction we perceive as inevitable, or *worried* by a future action or inaction we perceive as possible but avoidable, but it is hard to conceive of how we might *regret* what we have not yet done or forborne from.

We do sometimes say that we regret a dispositional cause of our own past action or inaction: perhaps I regret being so conservative with my investing since I might have otherwise invested in Facebook; perhaps you regret having such a short fuse since it caused you to snap at your friend the other day. I am inclined to understand these cases as regret of action coupled with a causal hypothesis: I regret not investing in Facebook during its initial public offering, and I believe that this failure is due to my conservatism; you regret snapping at your friend the other day, you believe that you did so because you have a short fuse. This understanding is suggested by the oddness of "regretting" a disposition that never manifests itself in regretted action or inaction: I would not say that I regret an overconservative financial temperament that I have always somehow overcome to make good investments; you probably would not say that you regret a short fuse that your background in stand-up comedy has always allowed you to successfully disguise as humor. Such dispositions may well *worry* us as possible causes of undesirable future action or inaction, but it seems malapropos to say that we *regret* them when no such action or inaction has yet transpired.

So far, we have characterized regret as an intentional state whose object is a past action or inaction. We can be more specific: I propose that regret is a *volitional* state. The object of this state is the opposite of whatever action or inaction one regrets—an action where one regrets

inaction, and inaction where one regrets an action. That is to say, its object is a *counterfactual past* action or inaction such that *my having done* or *forborne* is the complement of whatever I forbore from or did that I regret. That is to say, I regret that I φ-ed if and only if I will that I had not φ-ed. For example, where I regret that I did not invest in Facebook during its initial public offering, my volition is that I had done so; where your regret is that you snapped your friend, your volition is that you had *not* done so.

It may seem peculiar to think of regret as a volitional state because its object is in the past. Indeed, when we normally *will* something (for by 'volitional state' I mean a state of willing), what we will is a future action, a future forbearance from action, or in some cases a future state of affairs.[78] Perhaps it weighs against regret's volitional character that the central cases of volitional states are those with future objects. Weighing in favor of regret's volitional character is that volitional states are not just occurrent states accompanying action or inaction but also dispositional states to act or forbear when the appropriate circumstances arise: it can be my will to drink a Coke for the better part of a day even though I am not holding that prospect before my attention, provided that, other things being equal, I take advantage of my first opportunity to drink a Coke. I propose that regret is also a disposition to act or forbear: if you really regret snapping at your friend, you would now stop yourself *from having done* so if that were possible.

[78] When we will future states of affairs, we generally will whatever actions collectively conduce to that end—(rational) volition is structured by the norms of instrumental reason. This structure, in addition to obvious parallels in usage and meaning, raise the question of the relationship between volition and intention. While that relationship is not something I wish to consider here, it seems intuitive to say that if you intend φ then φ is your will. However, it seems stranger to speak of intending a counterfactual past action or inaction than to speak of willing a counterfactual past action or inaction. If, in fact, volitional states can have counterfactual past objects—as my discussion of regret would have it—then perhaps in this respect volition and intention are not coextensive.

Suppose, for instance, you snap at your friend over email. You instantly regret it. Moments later, you recall that Google, your email provider, offers an "undo send" feature that delays, for potential recall, the delivery of sent email for a predetermined amount of time. You have, in short, the opportunity to retroactively forbear from snapping at your friend. Other things being equal, you will naturally take advantage of this opportunity. In this way, regret and dispositional volitional states are functionally the same: both states are dispositions to act or forbear in the appropriate circumstances. Indeed, failing to be disposed to undo a regret seems to defeat regret's very attribution: all things being equal, if you were aware of "undo send" but declined to use it, then I think most people would be very hard pressed to consider you regretful of sending the email, at least for as long as "undo send" was available, you were aware of its availability, and you nonetheless did not use it.

Another way regret and volition might seem distinct is in their phenomenality. At the very least, regret is always unpleasant. Volition, on the other hand, is generally pleasant to the degree that success approaches, and unpleasant to the degree that it recedes. Thus, regret is at best an unpleasant *species* of volition. And with that conclusion I am in agreement: regret is a volitional state, but there are certainly volitional states other than regret. But against this conclusion one might adduce a further distinction between the unpleasantness of regret and the unpleasantness of unpleasant volition. For I take the unpleasantness of volition as success recedes to be that of fear, as when one is on the brink of losing a game. The unpleasantness of regret, on the other hand, need not involve fear, since the undesired outcome has already come to pass.

This phenomenal difference, however, is not by itself sufficient to show that regret and volition are distinct. Consider that the objects of fear are invariably in the future. Assuming that

the object of the fear characteristic of unpleasant volition is that volition's failure, then the unpleasant volition in question is a volition with a future object. But nothing about unpleasant volitions *in general* is revealed by this feature of unpleasant volitions with future objects, at least not without begging the question against the possibility of volitions with counterfactual past objects.

Because of the functional similarity between regret and dispositional volitional states, and because the phenomenal difference between regret and volitional states is not sufficient to show that the former is not a species of the latter, I conclude that regret is in fact a species of volitional state, namely, that it is a volition with a counterfactual past object. I realize that these considerations are not likely to convince people especially put off by the very notion of volitional states with counterfactual past objects. If the strangeness of volitional states with past objects is felt to be insurmountable, then I invite the reader to think of regret as a "volitionalistic" state—a state *functionally* similar to a volitional state, but accompanied (or perhaps partly constituted) by the belief that achieving its object was once possible but is no longer due to the passage of time.

B. Moral and nonmoral regret

I would like to distinguish *moral* regret from regret in general. After all, the presence of *nonmoral* regret would not rectify the problematic punishment of the martyr, the masochist, or the vengeful villain. Indeed, it could make their punishments' outcome more problematic still. The martyr, for example, might regret the murder because the ensuing conviction deprived him of the opportunity to take a grander, bloodier stand against the state and its postal cronies, perhaps by bombing a post office or assassinating the postmaster general. The masochist might regret the murder because it was foolishly committed in a jurisdiction known for its relatively humane penal

institutions. The vengeful villain, finally, might regret the murder because after reflecting on what he has done, he determines that he would have found more satisfaction in torturing and murdering his victim's family while forcing his otherwise unharmed victim to watch.

I call the regrets described in the previous paragraph *nonmoral* because they are not motivated by the balance of moral reasons[79]. The masochist and the vengeful villain regret the murders they committed because their own satisfaction was not thereby maximized, where the question of the moral value of their satisfaction does not arise, even if only to be taken for granted. Their regrets are nonmoral, then, because the most important moral reasons (those arguing against murdering his victim) are either missing or (in the case of the value of the perpetrators' satisfaction) present but not *qua* moral reasons. The martyr's case is somewhat different, as he regrets what he did because he could have struck a greater blow against tyranny—a moral evil. Now let us suppose that in this example, the postal employee murdered by the martyr is personally innocent of the tyranny that the government may or may not be perpetrating, at least aside from any moral complicity that may ultimately arise from drawing a government salary or delivering mail sent to and from government offices. Let us suppose further that prior to committing the murder, the martyr was cognizant of his victim's personal innocence—the victim's death was in the martyr's mind a means to a morally valuable end. So while the martyr's regret is perhaps not nonmoral in as total a sense as that of the vengeful villain or the masochist, it is nonmoral *to the extent that* the most important moral reasons are missing (again, those arguing against murdering the victim).

[79] I am using 'moral reason' in the sense of a *pro tanto* moral consideration in practical reasoning. By the 'balance of moral reasons', I mean whatever conclusion, or set of possible conclusions, following from practical reasoning that takes into account the relevant moral reasons.

III. Empathetic moral regret and successful punishment

I submit to the reader that a man not wanting to die is a compelling and generally decisive moral reason not to kill him. I submit further that this reason is salient to a person who has the life of another in his hands, and who is undeserving of and not about to deserve punishment for killing. Finally, I submit that an agent adequately responsive to moral reasons finds the conclusion that he ought not to kill someone compelling *because* (perhaps among other things) the person in question does not want to die—that is, without *requiring* the inferential mediation of a principle like the general impermissibility of killing people lacking death wishes. Bernard Williams has famously claimed that there is something wrong with a husband who is ultimately motivated by inferential mediation of a general principle to save his imperiled wife over an equally imperiled stranger.[80] If there is any truth to this observation, then there is, *a fortiori,* something wrong—certainly something insufficiently right—with someone who refrains from delivering the coup de grace to his bloodied, pleading victim, but not because his victim so obviously doesn't want to die, but rather because our assailant recalls that it is generally impermissible in these circumstances to finish people off. I call *volitional empathy* this phenomenon of taking and responding to someone else's will, by itself and without inferential mediation of a general principle, as a *pro tanto* reason to help or at least not to interfere with it, where the weight of this reason is proportional to the importance of the volition to its original subject. Volitional empathy is volitional in the sense that one person's volitional states in this manner affects the deliberative process behind another person's volitional states (rather than

[80] Williams 1982c, 18

involving emotional or doxastic states, for instance), and it is empathy in the sense that one person's will is prompting another person to will in harmony, so to speak.

Our perpetrators were clearly deficient in volitional empathy when they killed their victims—the fact that their victims did not want to die was either not salient to them as a reason not to murder, or it was salient but motivationally inert. The significance of this deficiency is that if it were to be belatedly remedied—if our perpetrators were to be moved belatedly by the wills of the people they murdered—then they would find themselves fervently but futilely willing that they not have done what they did. In other words, they would be filled with regret. I call *empathetic moral regret* the regret prompted in this way—by the belated and consequently futile response to a moral reason recognized in the unmediated way characteristic of volitional empathy.

Let us consider more closely what would happen were we to supply this deficiency. We might imagine our wrongdoers emerging from prison, or perhaps awaiting execution. Let us stipulate that they now experience volitional empathy with their victims: they take and respond to their victims' volitions as reasons to promote the volitions' objects, where the volitions in question include those our wrongdoers were aware of at the moment of their crimes, as well as those of which they were retrospectively informed in the course of their trial and punishment. Countless such volitions are associated with their victims' now annihilated projects: the promotion never to be earned, the charming house never to be bought, the Broadway show never to be seen with the spouse never again to be embraced, the children never to see grown, the estranged father with whom never to be reconciled. *Ex hypothesi*, the wrongdoers take each volition as a *pro tanto* reason to help or at least not to interfere. They are now belatedly moved by

the operation of volitional empathy to promote the volitions in question. But by their own actions, they made have it impossible to promote these volitions in any direct way. The wills they nullified have now at last become their own.

The existence of this state of empathetic moral regret is incompatible with the persistence of the hateful contempt of the vengeful villain, the delight of the masochist, and the self-righteousness of the martyr, as these attitudes each deny the reason-giving force of the victims' volitions. The contempt of the vengeful villain takes the volitions of his victims not as reasons for promotion or noninterference, but rather as urging their own obstruction—the very fact that the victim begs for his life argues that he should be deprived of it, or so the vengeful villain formerly reasoned. The masochist's deliberation was differently deficient: the victim's volitions were to be promoted or obstructed as a means to achieve the end of a pleasurable incarceration. The martyr, finally, neglected to take his victim's volitions as reasons at all, or took them as reasons, but with deliberative weight far short of what was appropriate given their importance to their original subject.

Empathetic moral regret clearly cannot coexist with such sentiments on pain of rational and motivational incoherence. The victory of empathetic moral regret quiets the intuition that our wrongdoers were not successfully punished. I think it does this because that intuition reflects the sense that the wrongdoers were not *responsive in the right way* to moral reasons: their *unresponsiveness* disposed them to *do* wrong, while the *manner* of their unresponsiveness disposed them *not to regret* wrongs already done. Empathetic moral regret is the consequence of, and thus evidence for, the existence of volitional empathy. Further discussion of volitional empathy is beyond the scope of this paper, but I offer that the need to rectify volitional empathy's

absence—which explains the absence of empathetic moral regret—is what *ultimately* explains the intuition that wrongdoers are not successfully punished unless they are made to regret their wrongs done.

IV. Sufficiency of regret for punishment

I conclude this paper with the observation that empathetic moral regret's role in punishment is more than merely a necessary condition for success. While it is not my ambition here to prove that empathetic moral regret is the sole purpose of punishment, I would like to gesture towards the conclusion that empathetic moral regret explains the effect that punishment should have on the wrongdoer. In favor of this point is that sometimes leniency or even forgiveness is warranted, and the presence of empathetic moral regret can explain why this is so. We are inclined to forgive people who have "suffered enough," as well as people who make a sincere effort to repair the harm they have caused. People sufficiently regretful satisfy both conditions—we might even say that they "self-punish." Regret under any conception is a form of suffering. We would expect the suffering of empathetic moral regret to be especially persistent, since it is a volition concerning the counterfactual past, and as such, it cannot be terminated by being accomplished. Nor can it be terminated by simply changing one's mind, since it is motivated not by one's self-chosen goals but by the appreciation of reasons which cannot be made to disappear at will. In addition, regretful people will be strongly disposed to repair whatever harm they caused to the degree that it is in their power to do so. This, again, is because regret is understood volitionally, and while they cannot achieve the object of their volitions—to undo their wrong—it would be only natural for them to seek what is next best, namely, to fix what they broke to the extent that doing so is in their power. Indeed, if they were not disposed to

try to repair their harms, we would take that as evidence that they did not really regret what they did.

Another reason we might think that inducing empathetic moral regret is the effect punishment should have on the wrongdoer is that it goes a long way towards explaining the other desiderata, which seem to describe different features of the regret to be induced by punishment, or how punishment is to induce it. The retributive desideratum of giving wrongdoers their moral deserts, where moral desert is suffering in a kind and proportion suited to the wrong, can be explained as seeking to induce in the wrongdoer the infringed volitions of the victim in kind and degree. When we are wronged, our retributive instincts do not simply demand that the wrongdoer be harmed, they demand that the wrongdoer *know how it feels*. Retaliating an inaccurate feeling, even if it is equally unpleasant, seems unsatisfactory. This way of analyzing retributivism has the additional benefit of explaining the mysterious "moral alchemy" Hart accuses retributivism of supposing "in which the combination of the two evils of moral wickedness and suffering are transformed into good."[81] For responsiveness to moral reasons is undeniably a moral good. Sometimes, it is unpleasant—excruciating, even—to confront and respond to reasons. But it is hard to see how anyone who values agency could maintain that the good of being an agent—insofar as agency requires responsiveness to reasons—is outweighed by whatever unpleasantness confronting those reasons involves.

The expressive desideratum, or at least the reprobative criteria (I take the expression of resentment to manifest fundamentally retributive concerns), seems to describe the consequences of regret if we understand the agent of the expression to be not the state but wrongdoers

[81] Hart 2008, 234

themselves. Regretful wrongdoers are disposed to willingly accuse themselves of their wrong, satisfying the primary reprobative function of expressing a stern judgment of disapproval. Regretful wrongdoers can in addition satisfy something analogous to *authoritative disavowal* by disavowing their own wrong as something they would never now do and could never in retrospect justify. They also discharge their special obligation to recognize their wrong as such, thereby expressing *symbolic non-acquiescence.* Regretful wrongdoers can *vindicate the law* by willingly resigning themselves to the mercy of the punishing authority, thereby recognizing the law's mandatory character. Finally, by assuming full responsibility for whatever it is they are fully responsible for, they *absolve others* of the same wrong.

Where the retributivist desideratum describes the nature and magnitude of punishment as a regret-inducing experience, and where the expressive desideratum describes the consequences of regret successfully induced, the educationist desideratum gets closest to identifying regret itself as an important goal of punishment in Morris's case, or its only goal in Hampton's. Hampton and Morris seem to diagnose in the wrongdoer a deficiency in moral knowledge, hence their conceiving of punishment as a kind of communication.[82] I grant that in a certain light, failure to appreciate reasons of a certain kind might be construed as or explained by a deficiency in knowledge. I maintain, however, that my way of looking at punishment's goal—inducing empathetic moral regret rather than communicating moral knowledge—is more illuminating: it is much easier to see how agency is respected when people are compelled to become responsive to moral reasons than when they are compelled to be taught moral knowledge.

[82] Morris 1981, 264; Hampton 1984, 216

Chapter 3: Normative Trust and the Confidence Trickster

Trust, as I understand it,[83] is not merely reliance: only trust can be betrayed, while mere reliance can only be disappointed. This important insight of betrayal as the hallmark of trust, introduced by Annette Baier in her formative discussion, has motivated the category of *will-based* trust theories, as Karen Jones has called them.[84] In order to distinguish reliance relationships where betrayal is possible from those where it is not, and in this way, to distinguish trust from mere reliance, will-based theories require trustors to rely specifically on trustees' good will, competence, and nothing else. Unfortunately, all will-based theories on offer are subject to Richard Holton's counterexample of the confidence trickster. Holton's trickster initially relies on a potential victim so that the victim will come to rely on him—a reliance the trickster intends to exploit at the victim's expense.[85] Intuitively, the trickster's reliance is not trust, and accordingly, it is susceptible only to disappointment and not to betrayal by the potential victim. And yet, the theories of Baier, Jones, Pamela Hieronymi, and Zac Cogley all count the trickster's reliance as

[83] 'Trust' is ambiguous in common use. For instance, had Admiral Nelson signaled to his fleet before Trafalgar, "England trusts that every man will do his duty" (close to the originally proposed "England confides…") he might have meant that England *hopes* that every man will do his duty, that England *believes* that every man will do his duty, that England *relies* on every man doing his duty, that England *demands* that every man do his duty, etc. I intend to discuss a sense of 'trust' where the trustor is *vulnerable to betrayal* by the trustee, and without prejudice to other senses of 'trust'.

[84] Baier 1986, 234–35. In addition to will-based theories there are *risk-assessment* views, which make no attempt to account for the intuition that trust can be betrayed, and which conceive of trust as depending indifferently on any consideration bearing on the trustee's likelihood of behaving as the trustor predicts, potentially including the effects of coercion, deception, and narrow self-interest. See, for instance, Dasgupta 2000; Hardin 2002. Hardin's concept of trust is slightly narrower than Dasgupta's. Hardin's encapsulated interest theory requires that the trustor believe that the trustee will come through because the relationship with the trustor is beneficial to him (Hardin 3). Dasgupta merely requires that the trustor believe that the trustee will do something, even if the trustee's reliability is unintentional (Dasgupta 55-56). I am inclined to see the will-based and risk-assessment camps more as interested in different concepts of trust than as disagreeing substantively.

[85] Holton 1994, 2

trust, vulnerable to betrayal. This is likely not by design for Baier and Jones, and certainly not for Hieronymi and Cogley, who both attempt to exclude the trickster case.

Cogley's solution is especially promising, benefiting from the important insight that trust has a normative component. As we shall see, however, it remains possible for a trickster to satisfy the conditions of Cogley's trust account. I am nonetheless convinced that Cogley was right to fix on the normativity of trust as key to solving the trickster problem. If only we had an account of trust specifically adapted to explaining trust's normative features, we might finally understand why it is that the trickster does not trust. And if ruling out the trickster case is a consequence of correctly accounting for these normative features, then solving the trickster problem can be used to guide the construction of a normatively focused account. In order to explore this possibility, the goal of this chapter is twofold: to solve the trickster problem, and to account for the normative features of trust. I proceed by constructing an account of trust designed to achieve both aims. If I succeed, my account of trust will not only have the advantage of explaining trust's normativity, it will also be extensionally superior to the will-based theories on offer, which fail to exclude the case of the trickster.

This paper divides into five sections. The first reviews the trickster problem and the solutions offered by Holton, Hieronymi, and Cogley. The second section considers the normative features of trust relationships, features to be explained by a normative account of trust. Importantly, it will turn out that trust is normative, but in a non-moral way. The third section considers a suitable form of non-moral normativity. The paper culminates in the fourth section, which proposes an account of trust that employs the form of normativity developed in the third section, that explains the normative features of trust relationships discussed in the second

section, and that solves the trickster problem discussed in the first section. The concluding section explores the significance of betrayal, which connects trust to regret and punishment.

I. Prior Solutions to the Trickster Problem

Holton proposes the trickster as a counterexample to the characteristic requirement of will-based trust theories, namely, that the trustor rely on the trustee's good will and competence. The trickster, after all, may rely on the victim's good will (and presumably, on her competence), yet intuitively, he does not trust her during his fraud.[86] The trickster's modus operandi is to rely on his victim to rely on him. He then exploits her reliance to his benefit and her detriment. For instance, suppose that a financial advisor approaches someone with an offer to invest her life savings. Predictably cautious, she decides to try him out by investing an insignificant sum with him, intending to withdraw her principal after several months—or whatever is left of it, if anything, should he turn out to be incompetent or a fraud. She tells him at the outset that she will invest her savings with him if he skillfully manages this smaller sum. Several months later, she withdraws her entire original investment as planned, and in the meantime, she has received several checks of market-beating returns. Concluding that the advisor is trustworthy, she invests a much larger sum with him. He turns out to be a confidence trickster after all, vanishing with her savings. In reality, the trickster did not bother to invest the initial small sum. His willingness to return it and to pay out "returns" were ruses to induce her to give him control over her savings. If

[86] Holton 1994, 2. In this same place, Holton also argues that good will is not sufficient for trust since a "member of an estranged couple between whom there is precious little goodwill can still trust the other to look after the children." The insufficiency of good will is undisputed by will-based theorists themselves, however, since belief in the trustee's competence is universally understood to be a necessary condition of the trustor's trust.

she had decided not to invest with him again, he would have lost his own money—perhaps a substantial amount for him.

Does the trickster trust his victim when, in order to profit and to recoup the phony returns, he relies on her to invest with him again? He does something outwardly resembling trust, certainly, but does he, in fact, *trust?* It does not seem right to say that he does: there is something intuitively disqualifying about his reliance's role as a means to defrauding his victim.

Can we explain this intuition by appealing to trust's hallmark, namely, the vulnerability to betrayal? Suppose that the trickster does trust his victim. In that case, she would betray him if she did not invest with him again after his excellent performance. Perhaps the situation struck her more and more as fishy, somehow, though she could not support her suspicion with reasons, nor was her suspicion itself an objectively reliable sign of untrustworthiness. But is it correct to characterize this decision as a betrayal? She does something outwardly resembling betrayal, especially if she led him to believe that she would hand over her savings should he do well with her trial investment. Despite this change of heart, however, there is something intuitively exculpating about her ignorance of his relevant intentions, since she would never have relied on him had she known them.

Holton's solution to the problem he raises is to propose that trust is reliance from the *participant stance*, a concept he borrows from P. F. Strawson, whose own term is the *participant attitude*. This is "the attitude (or range of attitudes) of involvement or participation in a human relationship," where one is disposed to feel *particular* participant attitudes[87] (the more specific

[87] Or *reactive attitudes* (7), or *participant reactive attitudes* (10), terms that Strawson appears to use interchangeably, and not to distinguish species or genus.

attitudes in the "range" of *the* attitude) towards another person, at least some of which are reactions to whether he or she manifests the "degree of goodwill or regard" that we "demand… on the part of those who stand in these [personal] relationships to us." The presence or absence of the expected "degree of goodwill or regard" activates a disposition towards gratitude or resentment independent of the good or harm incurred. For instance, "If someone treads on my hand accidentally, while trying to help me, the pain may be no less acute than if he treads on it in contemptuous disregard of my existence or with a malevolent wish to injure me. But I shall generally feel in the second case a kind and degree of resentment that I shall not feel in the first."[88]

Holton asserts that the trickster does not trust on account of relying on our good will without adopting the participant stance towards us—while "treat[ing] us simply as objects to be manipulated to his advantage."[89] Holton does little to elaborate on this claim, though he is clearly trying to articulate the intuition that the trickster is not altogether relating to us as one agent to another, but rather as an agent to a mere means.

Hieronymi offers an "admittedly very sketchy suggestion" to explain how, exactly, the trickster fails to take the participant stance towards his victim. For Hieronymi, if I were take the participant stance towards you, I would think of your intentions as products of reasons, and not merely as psychological facts useful for predicting your behavior. Now, I do not merely keep in mind the fact that your intentions issue from reasons, but rather, "I… allow your reasons to factor into my thinking and support my beliefs and decisions in something like the way my own

[88] Strawson 2008, 6–10
[89] Holton 1994, 5

[reasons] will." Were I to do this, "it seems right to say that I adopt the participant stance towards you."[90] The trickster, presumably, does not think of his victim's intentions in this way.

The sketchiest yet perhaps most crucial part of Hieronymi's "sketchy suggestion" about the participant stance concerns what she means by 'reason', and what, exactly, my relationship is to your "reasons" that makes them "something like" my own reasons. Two interpretations come to mind. The first is that your reasons are considerations that *justify* your intentions. In "something like" the way I am moved to find my own intentions justified by attending to my reasons, I am also moved to find your intentions justified by attending to your reasons. I set this interpretation aside because Hieronymi does not elaborate, nor is this interpretation what Cogley takes to be right when he disputes Hieronymi's solution to the trickster problem. I am sympathetic to this direction, however, as it suggests an intuition articulated in Karen Jones's generic concept of good will, Abraham Roth's concept of practical intimacy, and my own concept of volitional empathy, all of which I discuss below in my section entitled "Trust's Normativity."

The second interpretation of Hieronymi's participant stance suggestion is what Cogley takes her to be saying, and which he criticizes as failing to rule out the trickster's initial reliance as trust. On this interpretation, your "reasons" are the goals and the instrumental considerations salient to you that, together, *explain* your intentions. I take the participant stance towards you by keeping these explanatory factors in mind. These factors are "something like" my reasons, perhaps because, counterfactually, they *would* explain my having the intention that you *actually* have, were I to have your goals and were the same instrumental considerations salient to me. I might assume your point of view in this way to simulate your thinking, perhaps as if to predict your next

[90] Hieronymi 2008, 225–27

move in chess: were I to somehow play against someone exactly like myself, and were I to see exactly the same opportunities that you see, then the move I would make is the one I predict you will choose.

We recall that the point of Hieronymi's suggestion is to explain how the trickster does not take the participant stance towards his victim when he relies on her. If the suggestion succeeds, Hieronymi's participant stance view of trust could rule out the trickster's initial reliance, thereby solving the trickster problem. Against this solution, Cogley argues that the trickster can, in fact, take the participant stance towards his victim—the participant stance, at least, as understood according to the second interpretation of Hieronymi's suggestion. Cogley's first move is to argue that the participant stance view is too weak, counting too many cases of reliance as trust, including cases of mere reliance:

> [M]any cases of simply relying on another person will involve an implicit reference to that person's reasons, given that we are dealing with another agent. For example, I rely on other motorists to allow me to proceed first when I arrive at a 4-way stop before they have. If I consider why I rely on them it is because I have certain beliefs about their reasons: I believe they want to avoid accidents and being cited by the police, and that these reasons are sufficient to get them to follow most traffic laws. But simply bringing the content of the reasons of others to mind does not amount to trusting them. Nor would it be reasonable for me to feel betrayed if one of the other drivers entered the intersection before me. While anger or exasperation might be reasonable for me to feel toward the driver, it would not be appropriate to feel betrayed.[91]

[91] Cogley 2012, 38

Cogley applies the foregoing reasoning to show that Hieronymi's proposal would also count tricksters as trusting. Invoking the infamous Ponzi schemer Bernie Madoff, Cogley writes:

> When Madoff relies on his potential victim to go along with his plan, we can suppose that he has many true beliefs about the kinds of considerations that she tends to view as reasons. As a good-hearted person, she tends to support the plans of other people and she has even more of a soft spot for friends of her friends. Madoff knows that she tends to be frugal with her money, but also will take a chance from time to time on a more risky investment. Furthermore, he correctly surmises that she finds his current list of clients very impressive, which gives her confidence in his excellence as an investment advisor.[92]

In this example, Madoff does seem to bear in mind not only his victim's intentions, but also how these intentions are explained by the victim's goals and by the instrumental considerations salient to her. If bearing these factors in mind is all that it takes to assume the participant stance, then Madoff certainly assumes it.

Cogley's rebuttal of the participant stance view may not do justice to Strawson's original idea that both Hieronymi and Holton invoke. The participant stance, we recall, is supposed to be a disposition to react to your treatment of me with emotionally-valenced attitudes like gratitude or resentment. The disposition to these reactive attitudes must therefore be part of the participant stance, even if Hieronymi *also* thinks that a necessary condition of the participant stance is keeping someone's goals and instrumental considerations in mind. Indeed, the assumption behind the participant stance view may well be that the trickster is some kind of amoral sociopath, one emotionally detached from his victims in a way that normal people are

[92] Cogley 2012, 39

emotionally detached from inanimate objects. The emotional dispositions characteristic of Strawson's original idea would then be *of the essence* of the participant stance view. Thus, if Cogley means to show that the trickster can take the participant stance, he must show that the trickster can have a disposition to reactive attitudes towards his victims. It is not enough to show that the trickster can bear his victims' goals and instrumental considerations in mind.

Can the trickster be disposed to reactive attitudes towards his victims? Can he assume the participant stance in Strawson's original sense? In support of Cogley criticism, and against Holton and Hieronymi, I think that this is quite possible. Consider a trickster who is *not* an amoral sociopath. Jacques, we'll call him, is a successful accountant, a family man, and a pillar of his community. Unknown to nearly everyone, however, he is also deeply in debt to the mob, which has just delivered an ultimatum: he has one week to repay his debt before they collect their pound of flesh from his family. Jacques doesn't have the money or the assets, but he does have a number of wealthy friends. One by one, he appeals to them for help. He is vague about his situation, but his desperation is evident. His friends will be promptly repaid once he resolves his problem, he assures them, knowing full well that whatever money they lend him they will never see again. Privately, he is beside himself with guilt, but he would sooner bilk his own friends out of hundreds of thousands than abandon his family to a mafia hit squad.[93]

On Jacques's very last day to come up with the money, he approaches Grace, his best friend, whom he has known all his life. She listens to him with sympathy but finds herself very reluctant to help. Several years ago, you see, she lent an acquaintance a large sum of money that was never repaid, and since then she resolved to neither a borrower nor a lender be. Jacques

[93] Thanks to Jeff Helmreich for suggesting this example.

pleads with her, but at length she declines. In short, Jacques relied on Grace to rely on him, but her reliance was not forthcoming.

Pace Holton and Hieronymi, I offer this case as an example of how a trickster might take the participant stance, in the proper Strawsonian sense, towards his own victim. Under the circumstances, it seems not only possible but even likely that Jacques resents Grace. After all, he took her for his best friend, and in this time of his greatest need, he felt entitled to her help, even if it must be unwitting. It would be astonishing if Jacques's resentment and sense of entitlement were totally annulled by his consciousness of his own guilty intentions, or even by the real possibility that Grace would have given him the money he needed had he only been able to bring himself to tell her the truth.

The case of Jacques seems to support Cogley's rejection of the participant stance approach to solving the trickster problem, even if Cogley himself does not do justice to those solutions. As for Cogley's own solution, his strategy is to add a condition to the will-based analysis of trust, a condition involving a putatively normative relationship between the trustor and the trustee. Cogley's proposal is as follows:

> To trust someone is (i) to believe that because a person will be directly and favorably moved toward us we can count on her good will and competence governing our interactions in a particular domain and (ii) to believe that we are entitled to her good will because we are party to a normatively characterized relationship with that person.[94]

[94] Cogley 2012, 44

The trustor must not only believe in the trustee's good will, but believe that she is *entitled* to it. This entitlement constitutes a "norm of trust" that characterizes the relationship between the trustor and the trustee. The trustee betrays the trustor, endangering their relationship, when he violates the norm by failing to have the good will to which the trustor believes herself entitled. This violation occasions feelings like anger, disgust, or sadness—feelings that give betrayal its emotional color, though they are not themselves sufficient to distinguish betrayal from mere disappointment: "Feelings of betrayal tend to be more intense," reasons Cogley, "But the difference between feeling betrayed and feeling disappointed continues to assert itself even if we suppose that the intensity is equivalent across an episode of betrayal and disappointment. This is because feeling betrayed involves an implicit reference to the external situation—the breach of a relationship—that warrants the emotion."[95]

Cogley's general strategy for solving the trickster problem is to show that the trickster is not vulnerable to betrayal. Since trust requires that the trustor be vulnerable to betrayal, Cogley would thereby show that the trickster does not trust. How, then, is the trickster not vulnerable to betrayal? Returning to the example of Bernie Madoff, Cogley writes, "Madoff is not *entitled* to take his victim's good will toward him to get her to go along with his scheme. Consequently, Madoff's victim does not *owe* him her good will." And indeed, this plausibly explains why Madoff, as Cogley predicts, could not reasonably feel betrayed if a potential victim turned him down.[96]

Madoff could, however, feel betrayed. The feeling would be unreasonable, as many feelings are, but he could still have it, on account of having the false belief that he is entitled to

[95] Cogley 2012, 40–41. Cogley appeals to Margaret Gilbert's idea that relationships are constituted by norms—their "terms" (Cogley 2012, 46n16; see Gilbert 2006, 149-153)
[96] Cogley 2012, 36

his victims' good will, just as Jacques believes himself entitled to Grace's good will. Since Madoff would not *in fact* be entitled to his victim's good will, he is not *in fact* party to a normatively characterized relationship, or at least not one characterized by the norm of trust that Cogley posits. Cogley explicitly acknowledges such cases when he explains that his "account does not require that someone *actually* be a party to a relationship of the requisite kind," one whose constitutive normative character gives one an entitlement, "in order to count as trusting—it is sufficient for trust that someone *believes* himself to be a party to such a relationship." Cogley then gives the example of an unfaithful lover who really does *trust* his partner to be faithful despite having "no compunction about cheating" himself, and who is consequently vulnerable to feeling betrayed should his partner follow his faithless example. As Cogley explains:

> Given his own transgressions, his trusting would be unreasonable because his unwillingness to be faithful means he is not entitled to fidelity from his partner. He is not entitled to his partner being directly moved by the fact that he is counting on the partner not to cheat. My account of trust can correctly characterize this case as unreasonable, but extant, trust, because it does not require that the relationship *is* as the trusting one believes it to be—though reasonable trust does require this.[97]

The possibility of falsely believing oneself entitled to another's good will undercuts Cogley's solution to the trickster problem, since it means that the trickster can satisfy Cogley's conditions of trust. It is possible for Madoff (i) to believe that because his potential victim will be directly and favorably moved toward him, he can count on her good will and competence in the domain of writing cashing checks to well-regarded money managers sharing her cultural

[97] Cogley 2012, 44

background[98]; and, crucially, for Madoff (ii) to believe—however falsely—that he is entitled to her good will because he is party to a normatively characterized relationship with her.

The solution to Cogley's problem is straightforward: to avoid counting tricksters with false beliefs as trustors, we simply require trustors not only to *believe* but to *truly believe* themselves to be parties to normatively characterized relationships with their trustees. In other words, trustors must not only *believe* that they are entitled to their trustees' good will, they must *actually be* entitled to their trustees' good will.

II. The Normative Features of Trust Relationships

If we are going to say that in all cases of trust, a real entitlement exists—one that actually binds trustees—then we immediately face the following question: what is the nature of this normative relationship? Before giving an answer to this question, we should consider what an answer would look like—what features a normative theory of trust would explain. In this section, I collect these explananda.

It would be convenient if trust's normativity were simply moral normativity. After all, we can have false beliefs about our moral entitlements, and it is usually straightforwardly wrong, morally speaking, to betray someone's trust. Unfortunately, this route is problematic. As Philip Nickel writes:

> What does morality have to do with trust? Nothing much, one might say: the relationship of trust can hold between two apparently amoral people. For, first, it is possible to trust somebody to do what is morally bad; thieves and gangsters can be trusted rationally. Second, bad people, just like anybody else,

[98] Sales 2018

> can trust others. And third, people often use amoral criteria when they trust, in the sense that the grounds on which they draw the distinction between the trustworthy and the trustworthy are not moral grounds. Trust can be non-moral in its object, its hold, and its ground—even in all three at once. We can label this the Possibility of Amoral Trust.[99]

It appears that I must either show that trust is morally normative despite the Possibility of Amoral Trust, or I must describe a non-moral kind of practical normativity. I adopt the second strategy, which exchanges the paradox of amoral yet somehow morally normative trust for the question of what sort of practical normativity trust would have if not moral normativity. Let this, then, be the first normative feature of trust, namely, its non-moral character.

Trust's non-moral character is but one of several normative features that trust shares with Margaret Gilbert's concept of *joint commitment*, which I briefly sketch below. These features also include the special standing of the trustor to demand that the trustee perform and to rebuke the trustee who fails to perform. A further normative feature of trust, and one not shared with joint commitment, is the power of the trustor to unilaterally rescind a trust obligation.

Gilbert proposes joint commitment as the ground of the normative relations constituting shared agency, such as when you and I—when *we*—have lunch together, go for a walk together, jointly believe or decide something, form an agreement, and make and accept a promise, among many other cases. The ensuing normative state of affairs is one where we *owe* it to each other to get and to stay with the program—*our* program—so to speak. "Correlated with these obligations

[99] Nickel 2007, 309. Nickel cites Baier (1992, 110) and Hardin (2002, 75) for the first two points concerning the non-moral object and hold of trust. This possibility is reminiscent of Margaret Gilbert's discussion of immoral promissory obligations (Gilbert . Unlike Gilbert, who commits to the existence of such obligations as a manifestation of non-moral normativity, Nickel understands them as *incorrect ascriptions* of moral obligations. I will presently adopt Gilbert's strategy.

of the parties," Gilbert elaborates, "Are rights in the parties against one another: rights to actions that conform to the joint commitment. Correlated, again, with these obligations and rights is an important kind of standing or authority: the standing to demand conforming action and rebuke for non-conformity."[100] We should note that persons not party to a given joint commitment do not have the obligations, correlative rights, and correlative standing to demand and rebuke. For instance, I would not have standing to rebuke one stranger for breaking a promise to another, other things being equal, though the aggrieved stranger would have this standing. We should also take note of the non-moral character of joint commitment's normativity. The promise between the two strangers, for instance, might have been to do something immoral, and they would have been obligated accordingly. This is to say that the obligations of joint commitment are not moral obligations, nor is their existence subject to any moral constraints.[101]

We have already considered why trust normativity is non-moral, a conclusion motivated by the problem of reconciling a moral conception of trust with the Possibility of Amoral Trust. With respect to the special standing of the trustor, consider that there is nothing inappropriate, other things being equal, about a trustor mentioning a trust relationship while speaking with the trustee, just as there is nothing inappropriate, other things being equal, about a third party mentioning the trust relationship while speaking with the trustee. This would be unsurprising but for an almost inevitable crucial difference between the force of the trustor's mention and the force of the third party's mention. Consider that when speaking with a trustee, it is difficult for a trustor to mention their particular trust relationship—that the trustor trusts the trustee to see to some

[100] Gilbert 2015e, 8.
[101] Gilbert 2015g, 321–22.

particular thing—without implicitly demanding or rebuking.[102] Indeed, a very common trust demand takes the form of a simple report of the current existence of a trust relationship: "I'm trusting you." Almost any mention of the trust relationship by the trustor, however gentle or circumspect, is bound to have the air of demand. Similarly, a very common rebuke of betrayal takes the form of a simple report of the past existence of a trust relationship—"I trusted you." No mention of the betrayal itself is necessary. Almost any mention by the betrayed trustor to the erstwhile trustee of their breached trust relationship is accordingly bound to have the air of rebuke. I take this as reflecting the close conceptual connection between trust and the trustor's standing to demand and to rebuke—a connection so intimate that when speaking with a trustee, the trustor's mere mention of their trust relationship is enough for a demand or rebuke to be communicated.

A normative feature of trust not shared with joint commitment relationships is the power of the trustor to unilaterally rescind the trustee's obligation *qua* trustee.[103] This power exists because trust requires that the trustor rely on the trustee.[104] As a trustor can unilaterally cease to rely, she can unilaterally remove a necessary condition of the trust relationship grounding the trustee's obligation. This power seems to bear out in practice. Suppose that I trust you to drive me

[102] Perhaps the trustor might declare, "I trust you very much," or perhaps, "I trust you implicitly." But these statements do not so much refer to a trust relationship as they assert that the trustor finds the trustee trustworthy.
[103] Gilbert 2015b, 32. A joint commitment cannot be unilaterally rescinded by any single party.
[104] Trust requires reliance on my account, and also according many theorists across many camps, including Baier, Hieronymi, Dasgupta, and Holton. The role of reliance in Jones, Hardin, and Cogley is murkier, but trust for them still depends on the trustor's attitudes. While these may not be under the trustor's direct voluntary control, they nonetheless make trust (and any obligations it grounds) unilateral insofar as it depends on the trustor alone.

home from the airport tonight, and that you are consequently obligated to pick me up.[105] The day before I fly home, however, I decide to prolong my trip, and so I call you to say that you needn't pick me up after all. Your consent to being released is immaterial: even if, for some reason, you declare over the phone that you refuse to be released from your trust obligation, it is intuitively no longer in your power to betray me by failing to pick me up. It likewise seems clear that should I fly back on the original date after all—perhaps it proved too expensive to change my return flight—it would no longer be apt for me to rebuke you for betraying my trust by not picking me up. It seems, then, that whatever other obligations you may have concerning picking me up, your *trust* obligation was unilaterally terminated without your consent and, indeed, over your protests.[106]

III. Trust's Normativity

This section will discuss how trust can be normative without be *morally* normative. As mentioned earlier, the distinguishing feature of will-based theories of trust is that, other than competence, the trustor counts on something distinct from the trustee's interest.[107] This

[105] I use a pre-theoretical normative notion of trust here. It *at least* involves reliance, some kind of shared understanding between the trustee and the trustor, some kind of ensuing obligation binding the trustee to perform in accordance with the trustor's reliance, and the standing of the trustor to rebuke the trustee for betrayal when appropriate. I do not mean to suggest that the mere fact of my reliance obligates you to perform, whatever you may have to say about it.

[106] Perhaps you were looking forward to sharing a late dinner after picking me up. In fact, you've always picked me up from the airport, we've always had a meal together afterwards, and I've always scheduled my return flight for when you are available. You are now disappointed to hear me announce over the phone that I intend to return on a day when you are not free. Gilbert would probably analyze our little tradition as having constitutive joint commitment obligations that neither party can unilaterally rescind, and what is more, as involving an obligation on my part to arrive at the airport when you are available so that we can have dinner afterwards. Granting these obligations, the intuition that I can still unilaterally cancel your ability to betray my trust suggests that *trust* obligations are at least somewhat independent of the obligations of joint commitment.

[107] E.g., Baier 1986; Jones 1996; Hieronymi 2008; Cogley 2012. Hieronymi takes some inspiration from Holton 1994, who is not a good will theorist, denying as he does that good will is either necessary or sufficient for trust (2).

something is what will-based theorists call 'good will', and it is from a conception of good will that I plan to derive the normativity of trust.

What, exactly, do trust theorists themselves mean by 'good will'? Jones has criticized the literature's use of the term as "a meaningless catchall that merely reports the presence of some positive motive." That said, among these motives she discovers a "kind of unity": "If I have robust goodwill toward someone, of the kind found in friendship or good collegial relations, I will take the fact that they are counting on me to be a reason to act as I am being counted on in my motivationally efficacious deliberation."[108] Normativity is already implicit in this conception of good will since the agent takes the fact that others are counting on her as a *reason to act*.

For Jones, this generic good will does not appear to depend on one's social role, though one's social role may depend in part on the existence of good will. For instance, it is not because someone is a friend that one takes her reliance as a reason to act accordingly, though it is because (perhaps among other things) one takes her reliance as a reason to act accordingly that she is a friend.[109] This kind of role-independent responsiveness is similar to Abraham Roth's concept of *practical intimacy,* whereby one person can act directly on another person's intention, as Roth posits in his account of shared agency. The deliberative process Roth describes when one acts in this way is of particular interest:

> I need not have deliberated over, weighed, or recognized reasons for doing this or that. If anyone had to have done so, it was this other individual, the intender. I didn't decide or settle the issue of what to do regarding this matter; the other individual did that. Her decision does not merely provide a

[108] Jones 2012, 67
[109] 2012, 68. Compare Jones's concept of conscientiousness, which is role-dependent.

> strong *reason* for me to act as she intends for me to do, a reason that is supposed to enter into my deliberation about what to do. Rather, what we're envisioning is that I do not deliberate at all, or at least not enough to undermine this other individual's ability and authority to settle what it is that I'm to do.[110]

It is not entirely clear whether Roth sees practical intimacy as circumventing the agent's normal practical reasoning process (where "I do not deliberate at all"), or as submitting to that process a reason the agent takes as sufficient (perhaps I deliberate but "not enough" by considering all potential objections as answered without examining them myself).

In my own account of trust, 'good will' is cashed out in a way similar in some respects to Jones's generic concept of good will and to Roth's concept of practical intimacy. I call *volitional empathy* what it is, exactly, that I have in mind. The concept of volitional empathy shares with Jones's generic good will and Roth's practical intimacy the idea that one's actions can be connected to the practical reasoning of another agent in a way more direct than a mere coincidence of interests. Volitional empathy is more general than Jones's good will, however, in that one may volitionally empathize with a person who is not *already* relying on one (as when one is moved to offer unsolicited help). And unlike Roth's practical intimacy, volitional empathy involves volitions rather than intentions. While volitions and intentions are intimately related, they are importantly different, as I will discuss in my final section on the significance of betrayal.[111] In addition, volitional empathy does not short-circuit the agent's own practical reasoning, but rather submits to it a reason the agent takes as sufficient, or more precisely, as *controlling*. An

[110] Roth 2004, 384–85

[111] This distinction is especially important when relating trust to regret, which I conceive of involving past-oriented volitions. Past-oriented intentions may not be possible, as they are known to the agent to be futile.

agent's *controlling* reason is an internal reason psychologically analogous to a sufficient reason, in the sense that it would determine, on account of its importance to her, the outcome of her practical deliberation in conditions of adequate information. I distinguish controlling reasons from sufficient reasons to avoid the impression that controlling reasons must be sufficient in an externally normative sense.

Cutting to the chase,[112] an agent *G volitionally empathizes* with an agent *H,* with respect to an event *v,* if and only if:

(i) *H* wills that *G* sees to it that *v,*

(ii) (i) is an internal reason for *G* to will that *G* himself see to it that *v,* and

(iii) The weight of (i) in *G*'s economy of internal reasons is proportional to the urgency to *H* of *G* seeing to it that *v*.

By *volitional empathy* I mean all cases of volitional empathizing, and by *volitional empathy disposition* I mean an agent's disposition to volitionally empathize. A *concordant will* is a volition of the form acquired by volitional empathy—*G*'s volition that he himself see to it that *v* in (3).[113] *G*'s *volitional reason*$_{H}$ is *H*'s willing that *G* see to it that *v qua* consideration for *G* to will concordantly. A volitional reason may be *controlling* or *non-controlling* for *G* in the sense just described. Volitional reasons may also be *efficacious* or *inefficacious* for *G*, depending on whether

[112] See the introduction for a less abstract overview.

[113] Concordant wills are simply volitions with the same content belonging to different agents. They may issue from some process other than volitional empathy, like a prospective advantage to *A* or even sheer coincidence. When a distinction in time can be made, the concordant will is the volition occurring second. Thus, it is *A*'s will that is concordant (or concordant with *B*'s will), and not *B*'s will that is concordant (or concordant with *A*'s).

they *successfully move G* to will that he see to it that *v*—that is, depending on whether *G* is akratic.

I intend (i) and (ii) to capture a certain psychological and normative experience whereby another agent's goal-directed behavior seems to directly prompt one's help or one's noninterference. Declining this prompt is possible, but declining without justification warrants blame, and is in that sense and to that extent *wrong*. In more psychological terms, volitional empathy is the normatively laden attraction to helping and aversion to ruining what one perceives as a goal directed effort, together with the action motivated and warranted by this normatively involved attraction and aversion.

I wish to stress three points. First, the reasons of volitional empathy are *internal* reasons in Bernard Williams's sense, which I take to be considerations playing a reason-giving role in an agent's practical reasoning given that agent's set of motivations, as well as considerations that *would* play this role were the agent aware of them.[114] I set aside, as much as possible, the question of how volitional reasons are related to motivation-independent "external" reasons; as well as how volitional reasons are related to other internal reasons, except that I take for granted that it is psychologically difficult and uncomfortable to hold incompatible internal reasons where one is aware of this incompatibility, and that we are disposed by this and perhaps by other factors to resolve incoherencies of this kind.

Second, the volitional empathy disposition does not require indiscriminate volitional empathizing. Suppose, for instance, that one spots *M* mugging *P*. Any volitional reasons$_M$ may be

[114] Williams 1982a

swamped by more salient volitional reasons$_P$, perhaps to the point where considerations in favor of the mugger don't even come to mind. Alternatively, volitional reasons$_M$ might be swamped or defeated solely by nonvolitional considerations such as a general moral conviction against assaulting and robbing people, or by such considerations in combination with volitional reasons$_P$. That said, I don't rule out the possibility that volitional reasons$_M$ might be controlling for an agent wearing the right sort of blinders, or at least salient enough to be noticed—if only to be overruled—by someone not thus equipped, as one might come to root for a Michael Corleone or a Tony Soprano.[115]

Third, the reason-taking of volitional empathy does not involve the explicit or implicit mediation of a general normative principle covering the case under consideration.[116] For example, volitional empathy might involve the sight of a stranger trying to get up from a fall, where that sight *by itself* prompts one to help her. To feel prompted in this way, one does not need to consult a general principle requiring one to help others in cases like this. Rather, the stranger's very effort to get up seems to require that one assist, or at least that one not interfere. Whether one is prompted to assist or merely not to interfere may depend on the circumstances of the case. For instance, if the person sprawled on the pavement looks at you plaintively as she struggles to rise to her feet, then it will likely seem that you are required to assist, and not merely to refrain from kicking her back to the ground.

---

[115] Blinders might include racial prejudice or some other syndrome of beliefs or dispositions that arbitrarily denies or diminishes someone as a possible subject of volitions. For instance, Hitler's description of Jews as a "racial tuberculosis" ("Adolf Hitler's First Anti-Semitic Writing" n.d.).

[116] Although volitional empathy might supply intuitions one might take as evidence for a general moral principle. But volitional empathy would precede rather than be preceded by such a principle in the order of discovery.

IV. Normative Trust

In this section I develop my normative account of trust, where the normativity I have in mind is the non-moral variety described in the previous section. This account aims to solve the trickster problem, to explain the power of the trustor to unilaterally rescind the trustee's obligation *qua* trustee, and to explain the special standing of the trustor to demand and rebuke.

With one elaboration, I take for granted the concepts of *competence* and *reliance* when I say that an agent *R* may *rely* on another agent *E* to see to it that *v*, and where *E,* in turn, can be *competent* to see to that *v*. The elaboration is that *R*'s reliance on *E* to see to it that *v* is constituted, at least in part, by *R*'s volition that *E* see to it that *v*. I also make use of the concept of good will, which I understand in terms of volitional empathy as described earlier.

With these concepts in place, we may say what makes someone *trustworthy*. Namely, *E*, a potential trustee, is trustworthy with respect to *R,* a potential trustor, to see to it that *v,* an event, if and only if:

E is competent to see to it that *v,* and
E bears *R* good will with respect to v: if *R* relies on *E* to see to it that *v,* then <u>*E*</u> will have a controlling volitional reason$_R$ to see to it that *v*.

I mention trustworthiness first because will-based trust involves reliance on the *trust-warranting* (rather than *mere reliance warranting*) considerations bearing on a trustee's

reliability.[117] These considerations deserve to be understood as the conditions of trustworthiness on a will-based trust theory, since they alone can make someone worthy of another's trust.[118]

Trustworthiness is not itself trust as it cannot be betrayed: however surprised or disappointed we might be, we would not say that we are betrayed if someone on whom we are not currently relying, but whom we considered trustworthy, turns out to be a fraud and a pathological liar. The possibility of betrayal seems to require actual reliance—in particular, the reliance specifically characteristic of trust, which takes into account only the trustworthiness of the trustee as it appears to the trustor (and thus allows the possibility of trusting the untrustworthy).

Trust reliance (T): *R* relies on *E* to see to it that *v* only because *R* believes that *E* is trustworthy with respect to *R* and *v*.

If offered on its own as a will-based theory of trust, trust reliance would be hard to pick out from a lineup of other will-based theories, and in fact, closely approaches the first condition of Cogley's account of trust—a condition that is, in essence, Jones's account of trust with a friendly revision replacing Jones's attitude of optimism with Cogley's belief attitude.[119]

---

[117] Hieronymi 2008, 218–24

[118] All will-based theories aside from Baier's require the trustor to bear an attitude towards the trustee's competence and good will that disposes the trustor to rely on him. The theories of Hieronymi and Cogley require the attitude of belief, while Jones requires a weaker "attitude of optimism" (1996, 4) that Cogley critiques as inadequate to distinguish trust from hope and as failing to make the attitude *warrant* trust reliance (2012, 35). Baier does not specify an attitude, and specifically rules out belief, which she considers incompatible with unconscious trust, though she does not seem to consider the possibility of dispositional belief. It is, in any event, unclear how Baier would analyze the ill-advised trust—but trust nonetheless—of someone whose good will or competence is lacking. Interposing an attitude, such as belief, resolve this difficulty.

[119] Cogley 2012, 35–36

A trickster, however, can satisfy the trust reliance condition. Trust reliance is therefore at best one of several conditions of trust. We might strengthen it with another condition, reminiscent of Cogley, and informed by the normative insight:

> **Control belief (B_RC):** *R* believes that if T, then *E* has a controlling volitional reason$_R$ to see to it that *v*.

B_RC is similar to the normative condition proposed by Cogley, though instead of the trustor directly imputing to himself an entitlement, B_RC has the trustor impute to the trustee a controlling internal reason as a consequence of the trustor's trust reliance. And like Cogley's normative condition, B_RC goes some way towards explaining the trustor's feeling of betrayal as a response to being denied something to which the trustor believes she is entitled.

Unfortunately, B_RC does not exclude all tricksters. As with Cogley's condition, B_RC can be satisfied by tricksters who genuinely but falsely believe that the trustee is obligated to perform, as Jacques feels about Grace. As we discussed earlier at length, we must take the normativity of trust outside the trustor's belief context. Doing so yields the following condition:

> **Control (C):** if *R* trustingly relies on *E* to see to it that *v*, then *E* has a controlling volitional reason$_R$ to see to it that *v*.

The trustee is now normatively bound by trust, since it is not merely in *R*'s mind that *E* ought to follow through. If anything, the urgency of following through is in *E*'s own mind, since it is on pain of rational incoherence relative to *E*'s own motivations that *E* ought to see to it that *v*. As earlier discussed, this is because volitional reasons are internal reasons arising from volitional empathy with *R*, as prompted by *R*'s reliance on *E* to see to it that *v*. It is important that these reasons arise from the volitional empathy involved in *E*'s good will towards *R*: it would not do for

E's reason to be independent of *R*'s trusting reliance. In that case, it would be difficult to understand betrayal as a violation of trust's normativity, rather than as a violation of normativity independent of trust (e.g., as a moral wrong).

Using C, we finally exclude from cases of trust the trickster who falsely believes that his victim is normatively bound to perform. The trickster can still trustingly rely on his victim, but this reliance could no longer plausibly give his victim a controlling volitional reason to perform: the fact of the trickster's trust reliance would be outweighed by the danger to the victim arising from the trickster's intention to defraud. The victim may well be unaware of this danger, but its status as an internal reason for her is unaffected by her ignorance, following Williams's conception of internal reasons. According to Williams, this is because internal reasons must, as reasons, not only explain but also rationalize an agent's actions relative to her motivations. The trickster's intention to defraud would thus prevent his initial trust reliance from rationalizing the victim's decision to participate in his scheme. This is even if the victim's participation is *explained* by the trickster's trust reliance and by the victim's ignorance of his intention to defraud her.[120]

C cannot complete my account of trust, however. Intuitively, trust requires some kind of mutual understanding between the trustor and trustee, an understanding that the trustor's trust reliance and the trustee's responsiveness to such reliance do not quite capture, and which is required to explain the trustor's special standing to demand and rebuke the trustee.

To explain how trust on my analysis produces standing, it will be very helpful to conceive of standing using the same conceptual resources employed in explaining trust, namely, the

[120] Williams 1982a, 102–3.

apparatus of volitional reasons. From this perspective, standing is a kind of psychological leverage that a trustor has over a trustee, which the trustor has to the extent that the trustee is averse to the rational incoherence of rebelling against one of his own internal reasons—a certain internal reason, to be specific, with which the trustor is uniquely able to confront him. As discussed earlier, the trustee's reason, *qua* trustee, is the trustor's volition. This volition, for the trustee, is an internal reason for him to will concordantly. Now recall that when the trustor says, "I'm trusting you," her utterance has a different force than when a third party says, "*R* is trusting you." I propose that this is because where the third party *refers* to the trustor's volition, the trustor *presents* it by directly manifesting her volition as a *conspicuous trying*. That is to say, the trustor is not merely referring to her volition, but enacting it by trying, by means of her utterance, to carry it out—to get *E* to see to it that *v*. Nor is there anything discreet about this trying: not only does *R* try to get *E* to comply by means of mentioning the reason for his compliance, but *R* intends for *E* to notice that she is trying. It is no accident that *R* turns the screws on *E*'s conscience. In this way, *R confronts* E with his own internal reason, which also happens to be *R*'s volition.

The trustor's standing is special, not being shared with third parties, which we can see by considering a revealing exception, where she endorses a third party as her *representative,* speaking on her behalf. In this case, the representative's utterance mentioning the trust relation to the trustee is itself the trustor's visible trying—it is the representative's effort on the trustor's behalf that manifests her volition. This exception is illuminating because a third party who does *not* speak on the trustor's behalf may still speak *as if* on the trustor's behalf. But in that case, we expect that the trustee may reply, and aptly, "It's none of your business."[121] On my analysis, this

[121] See Gilbert 2015f, 288.

rebuffs the third party for the false pretense of enacting the trustor's will when in fact he speaks for the sake of his own distinct interest. Compare this with the case where the trustee accepts that the third party speaks for the trustor, but refuses to speak with anyone but the trustor herself about their trust relationship. "It's none of your business" seems no longer so apt, as we would expect on my analysis, seeing that the trustee here does *not* believe that the representative's agent is pursuing his own interest distinct from the trustor's claim. Indeed, we would expect the trustee to say something more like, "You can tell *R* that she can talk to me herself," perhaps adding, "*R* shouldn't have involved you," which declines the conversation but accepts that the third party speaks for the trustor.

I propose two further conditions of trust, KT and KC, that include and expand on the foregoing conditions T and C. These more comprehensive conditions complete my account of trust by explain how trust relationships give rise to the trustor's special standing to demand and to rebuke. KT and KC, which entail T and C, suffice for a full statement of my theory.

Known Trust Reliance (KT):

1) *E* knows that T, namely, that *R* relies on *E* to see to it that *v* only because *R* believes that *E* is trustworthy with respect to *R* and *v*, *and*
2) *R* knows (1).

Known Control (KC):

1) *E* knows that C, namely, that if *R* trustingly relies on *E* to see to it that *v*, then *E* has a controlling volitional reason$_R$ to see to it that *v*, *and*
2) *R* knows (3).

KT1 and KC1 function together to compel *E*, on pain of conscious rational incoherence, to see to it that *v*. Without KT1, *E* does not feel the pressure of rational incoherence because the antecedent of the conditional in KC1 does not obviously obtain. Even if it *does it fact* obtain—even if, for instance, both T and C obtain, *E* would not feel compelled since *E* does not *know* that T and C obtain. That said, *E* might *learn* that T and C obtain and consequently feel compelled, distinguishing that case from where T or C does not, in fact, obtain.

Knowledge of KT1 and KC1 would together allow *anyone to soundly argue to anyone* that *E* knows he should see to it that *v*: *E* knows that P, E knows that if P then Q, therefore, *E* knows that Q, on the assumption that E's knowledge is closed under known and straightforward deduction. In contrast, *R*'s knowledge of KT1 and KC1, required by KT2 and KC2, allow something in addition, and they allow it to *R* alone: *R* may not only to *soundly argue to anyone* but also to *demand of* E *in particular* that *E* see to it that *v*. This is because *R* alone can manifest her volition by asserting or implying that *E well knows that T and C*, making salient and occurrent to *E* the rational pressure to see to it that *v*.[122]

V. The Significance of Betrayal

I conclude by exploring the significance of betrayal, which connects trust to the phenomena of regret and punishment by way of the betrayed trustor's standing to rebuke. As earlier discussed, when *R* trusts *E, R* is able compel *E* to will that *v* on pain of rational incoherence.

[122] Gilbert has argued that the standing to demand an action in enforcement of a right requires a *joint commitment* among the parties involved. I touch on joint commitment on p. 13 and on pp. 85-88). Many of the rights Gilbert has in mind—moral and legal rights, in particular—are rights in a much more robust sense than the normative relations discussed in this chapter (Gilbert 2018). While I do believe that these relations qualify as rights in a structural sense, much needs to be done to relate them to the more robust, externally normative relations that Gilbert's account underwrites (see note to p. 111).

That is to say, while *E* might conceivably decline to will that *v,* to do so would be for *E* to rebel against his own controlling volitional reason$_R$, which is an internal reason of his that he acquires by means of his volitional empathy disposition. *R*'s ability to compel *E* in this way is *R*'s standing to demand under a psychological description. *R*'s ability to compel *E*'s will need not end if *E* betrays her—if *E* fails to will concordantly in some irrevocable way, such as by disclosing a secret that *R* told him in the strictest confidence. Her ability to compel *E*'s will under these new circumstances—her standing now to rebuke instead of demand—may persist so long as *R* still wills that *E* see to it that *v,* and so long as *E* still takes *R*'s volition as a reason to will concordantly.[123]

R's volition that *E* not have done what he has done already has an air of paradox about it. How can *R* will that *E not have done* what he has done already? How can she will that he *not have revealed* the secret he disclosed? I consider this question at greater length in my first chapter, where I proposed to explain regret as just such a past-oriented volition that can no longer be carried out. I might currently regret disclosing a friend's secret, which on my analysis is a volition I *now* have that I not *then* have disclosed his secret. This is not to say that I *now intend* to have *then done otherwise*—I do not *intend* what I believe impossible, namely, changing the past, though I do *will* it.

While a full examination of the relationship between intention and volition is well beyond the scope of this paper, I will observe that they are not equivalent, lest we be tempted to equate intending what we believe impossible with willing what we believe impossible. Volition is no doubt a necessary condition of intention, but it is not controlling. For instance, despite fully

[123] This is consistent with Gilbert's description of rebukes as "after-the-fact" demands (Gilbert 2015d, 417).

believing that it is impossible for me to beat Tiger Woods in a fair round of golf, I might nonetheless have the volition to do so. What does this mean? I might spend every waking hour and every cent in my bank account training to become the best golfer I can be, I might move heaven and earth to arrange to somehow play a round with him, and on the appointed day, I might make my utmost effort on the golf course. Under such circumstances, it seems strange to say that it is not my volition—my *will*—that I beat Tiger Woods. Yet if someone asks, "Do you really intend to beat Tiger Woods at golf," it would also be strange for me to say that indeed I do. At most, I might say that I intend to *try*. It is precisely this conative posture towards φ, this *trying* to φ, disposing one to *intend* to φ, were one to believe that φ-ing were possible, that I consider central to volition (perhaps among other conditions). This posture explains the content and unpleasant affective character of regret: one regrets not achieving what one would intend were achieving it only possible, and one feels frustrated and undermined as if struggling against the inevitable. This is precisely the attitude that *R* can compel *E* to experience should *E* betray her. So while *R* can no longer compel *E* to *will and consequently to intend* v, she can still compel *E* to *will* v *and thereby* to *regret not* v-*ing*.

In my second chapter, I proposed inducing regret as a moral goal of punishment. After all, punishment is intuitively imperfect whenever punishment subjects do not afterwards regret what they have been punished for. As I argued in that chapter, this intuition is consistent with, if not altogether explained by, the prevailing retributive, expressive, and educationist theories of punishment where the effect of the punishment on the punishment subject is a principal concern.

R rebuking *E* is in this way analogous to punishment. Punishment might even *be* a form of rebuke, perhaps given an appropriate background of social conditions.[124] In addition, the authority to punish, widely held to be a conceptual component of punishment, finds a ready analogy in the trustor's standing to rebuke a trustee for betrayal.[125] This analogy suggests that the authority to punish is found in a punishment agent's psychological leverage over a punishment subject's internal reasons. While punishment agents are probably, for the most part, third parties to the offenses punished, their standing can be supported using the concept of third-party agency discussed in the previous section. A final analogy is between the power of a punishment agent to forgive an offense and the power of a trustor to unilaterally rescind a trustee's obligation: when betrayal has occurred, the trustor's rescission removes the ground of the trustee's regret; likewise, when an offense is committed, the punishment agent's pardon removes the warrant for the offender's punishment.

[124] Gilbert considers rebukes to be a "mild form" of punishment (Gilbert 2015c). Feinberg's position that punishment has a "reprobative" function also suggests this idea (Feinberg 1965, 403-408).
[125] See, for instance, Flew 1954, 294; Feinberg 1965, 397; Hampton 1984, 225; Morris 1981, 264.

Chapter 4: Why Should We Rebuke?

Rebukes are intimately related to many normative phenomena, including trust, punishment, regret, blame, and shared agency. Despite these links, they have not yet attracted substantial attention, attention which they warrant on account of their central role in normative life, and on account of being puzzling. Consider the following conditional. *If* it is appropriate for one agent to insist to another that he conform to a given interpersonal norm when she doubts him, *then* it is appropriate for her to insist to him that he *should have* obeyed the norm when she knows he hasn't. For instance, if it is appropriate for you to insist that I adhere to our contract, to pay for items taken from your store, or to return your car undamaged—you have your doubts—then, when I fail to adhere to the contract, pay for the items, or return your car undamaged, it is appropriate for you insist that I *should have*. The point of the prospective cases of insisting—of *demands*—seems clear: to the addressee to do as he should. The point of the retrospective cases—of *rebukes*—is not so clear. Here I borrow 'demand' and 'rebuke' back into their pre-theoretical sense from Margaret Gilbert, for whom these words are technical terms from her theory of shared agency. I also borrow the insight, which illuminates the conditional above, that rebukes are "after-the-fact" demands[126] in some sense.

Even though it is not clear what rebukes accomplish, or even whether it is reasonable to issue them, is it clear that they are somehow appropriate. Perhaps rebukes are essentially an expressive variety of punishment communicating disapproval and resentment.[127] Perhaps rebukes

[126] Gilbert 2018, 63.
[127] Feinberg 1965, 403

just make people feel better. Perhaps they are uselessly, counterproductively vindictive. Perhaps different kinds of rebukes serve different purposes, some more reasonable than others.

I suspect that each of these answers is sometimes correct. Perhaps each is correct for a different species of rebuke. While it is my hope that this discussion illuminates every species to some extent, I confine my attention to one species of central importance: the rebuke of the betrayal of trust. Where the distinction matters, I call these *fiduciary rebukes*, though for brevity, I will usually just call them *rebukes* with the understanding that fiduciary rebukes are what I have in mind. When we attend to fiduciary rebukes in particular—that a trustee should have kept one's trust—their puzzling character is not lost. The horse has left the barn. How does it make sense to close the door as if his escape could still be barred?

By applying and elaborating on the account of trust developed in the previous chapter, I aim in this chapter to construct answers to a pair of closely related questions: what fiduciary rebukes do and whether they are reasonable. In answering these questions, I will take as seriously as possible the insight that rebukes are after-the-fact demands. The answer I develop, very roughly, is that the function of a fiduciary rebuke is to compel the trustee, on pain of rational incoherence, to repair the harm caused by the betrayal of trust—the peculiarly fiduciary violation—both to the trustor's ends, and to the trustor's relationship with the trustee.

I begin by briefly recapitulating my account of trust and identifying the relevant concept of betrayal. I then explain the prospective responses to betrayal available to the trustor, which include unilaterally rescinding her reliance volition, which terminates the trust relationship and contributes to estranging her from the trustee; and alternatively, demanding performance, which preserves the relationship instead. Finally, I consider the retrospective responses to betrayals that

have already occurred. To explain how rebukes are, indeed, after-the-fact demands in every sense, I consider past-oriented volitions, along with what I take to be two principles of rational coherence.

I. Trust and Betrayal

'Trust'

In this section, I summarize the features of the trust account developed in the last chapter that shed light on the function and reasonableness of fiduciary rebukes. Let's begin by settling the relevant sense of 'trust', since we sometimes use the word broadly to refer to reliance relationships of every kind. In some of these relationships, which I call relationships of *mere reliance,* the reliant agent counts primarily on her prediction about the behavior of the agent relied on. For example, an extortionist might "trust" his victim not to contact the authorities, predicting that he would be too humiliated or afraid to do so. Likewise, Kant's neighbors might "trust" him to appear outside their windows at the same time every day, predicting that nothing could interfere with such a fastidious man's long-established habit.

The distinguishing feature of mere reliance relationships is that the agent relied on cannot betray the reliant agent by not behaving as predicted. Intuitively, if the extortionist's victim does contact the authorities, then he does not thereby betray the extortionist; likewise, Kant does not betray his neighbors should he decide one day to take his walk at a different time. This intuition appears to be unaffected by how much the reliant agent has at stake. For instance, we would not say that the victim betrays an extortionist who risks a long prison sentence, but not one who risks being chided by her schoolteacher.

Closely connected to the impossibility of betraying mere reliance is the inaptness of rebuke. It would be absurd for the extortionist to "rebuke" his victim, or for the neighbors to "rebuke" Kant. This is because betrayal is the ground for rebuke in the context of a trust relationship, but the "trust" of mere reliance is not the right kind of trust.

I simply call *trust relationships* the reliance relationships in which betrayal is possible. These have been discussed at length by "will-based" trust theorists,[128] as Karen Jones has called them, since these philosophers seek to distinguish trust as a reliance relationship where the trustor's reliance is grounded in her belief in the trustor's good will towards her, along with her belief in his competence to do as trusted. [129] In the case of *mere* reliance relationships, reliance is grounded indifferently on anything relevant to predicting behavior of the agent relied on, whether or not those factors include the trustee's good will. For example, if you count on me to know German because you know that I have a German last name, then you merely rely on me (regrettably, to your disadvantage).

II. Trustworthiness, Good Will, and Trust

My account of trust is a will-based theory, requiring that the trustor's reliance be grounded in her belief in the trustee's good will and competence. Since the trustee's good will and his competence qualify him as a candidate for her trust, I refer to these conditions together as his *trustworthiness.* That is to say:

An agent *E* is *trustworthy* to an agent *R* with respect to an event *v* if and only if:

(a) *E* is competent to see to it that *v*.

[128] See, e.g., Baier 1986; Jones 1996; Hieronymi 2008; Cogley 2012
[129] Jones 1999, 68.

(b) *E* bears *R* good will with respect to *v*.

Will-based accounts of trust typically hold that a trustor trusts a trustee when she relies on him specifically because she believes that he satisfies these conditions. My account builds on this foundation by elaborating considerably on good will, a phenomenon that will-based trust theorists are generally content to take for granted.[130] My account also adds knowledge conditions to the effect that the trustor and the trustee see eye to eye about the trustor's reliance and the trustee's good will. All of these elements contribute to explaining how a trust relationship implicates interpersonal norms.[131]

In the course of explaining how trust is normative, my concept of good will also explains how a trustor can make fiduciary demands and rebukes. Very generally, the trustee's good will allows his own internal reasons to be used by the trustor to bind him to keep her trust. It is in this way that good will generates an interpersonal trust norm. The trustor can use the same process to increase the weight of the reasons by which the trustee is bound already, and it is in this way that good will empowers the trustor to demand and to rebuke.

[130] Jones 2012, 67

[131] In addition, they also allow my account to rule out Richard Holton's confidence trickster counterexample, which has been a troublesome problem for other will-based accounts of trust (Holton 1994, 64–65). The confidence trickster relies on a mark to rely on her so that she can ultimately defraud him. Intuitively, the trickster's reliance is not a case of trust, even though her reliance may very well be grounded in her belief that the mark bears her good will and is competent to write a big check. Consequently, reliance grounded in the belief that someone is competent and trustworthy is not sufficient to constitute trust. My account rules out the trickster case by showing that the confidence trickster's initial reliance does not, in fact, give the mark a sufficient internal reason to do as trusted, even if she thinks otherwise because she is ignorant of his fraudulent intentions. The case is similar to Williams's example of the agent who wants a gin and tonic, believes a substance is gin, but does not, in fact, have an internal reason to drink it mixed with tonic as the substance is really gasoline (Williams 102-103).

This mechanism for good will involves *volitional empathy,* namely, the disposition to take the volition of another agent as one's own internal reason. In the case of trust, the trustor's volition is that trustee see to the event that she, the trustor, is counting on, which the trustee takes as an internal reason that he, the trustee, see to that event. For instance, if I am trusting you to keep a secret, then you *volitionally empathize* with me with respect to keeping the secret provided that you take my volition that you keep it as your own internal reason to do so. The weight of that reason and the urgency of the volition are proportional: introducing the secret by saying, "Just between you and me," gives you reason to keep it, but less reason than were I to take you aside, look you in the eye, and say with a lowered voice and a hand on both your shoulders, "It is very, very important to me that you keep this close to the vest. I cannot stress to you how essential it is that no one ever find out." In general, *E volitionally empathizes* with *R* with respect to an action *v* if and only if:

(iv) *R* wills that *E* sees to it that *v*

(v) (a) is an internal reason for *E* to see to it that *v*

(vi) The weight of (a) in *E*'s economy of internal reasons is proportional to the urgency to *R* of *E* seeing to it that *v*

The trustee's internal reason to have this will—that reason just being the trustor's volition taken by the trustee as an internal reason—just is the trustee's *volitional reason*. I call an internal reason *controlling* for *E* if it predicts the volition that, given enough relevant information (a state which may be the rule rather than the exception), would leave him feeling less at odds with his own ends—less as though he foreseeably undermined what he most values—than if he did not will accordingly. The controlling reason predicts this volition even if the option of not willing

accordingly is supported by contrary, non-controlling internal reasons that *E* happens to have at the same time. This is simply to say that contrary internal reasons can both be inputs into one's practical reasoning, a process that nonetheless may decide in favor of what one reason supports over what a conflicting one does.[132] Where a trustee has a controlling volitional reason, and where this reason issues in a volition that he see to it that *v,* this volition is his *concordant will*—concordant, specifically, with the trustor's will, with which it shares its content.[133]

I can now restate the good will condition of trustworthiness in terms of a controlling volitional reason. Trustee *E* bears trustor *R* good will with respect to an event *v* if and only if the following conditional obtains:

> If *R* wills that *E* see to it that *v,* then <u>*E*</u> has a controlling volitional reason$_R$ to see to it that *v*.

With this concept of good will in mind, here are the conditions of trust.[134] *R* trusts *E* to see to it that *v* provided that:

(1) *R* believes that *E* is trustworthy with respect to *R* and *v,* namely, that:

(a) *E* is competent to see to it that *v*

[132] This is not to say that there is no possible contrary internal reason, or combination of them, that could become controlling later, where the original controlling reason recedes in importance and becomes non-controlling. People have changes of heart, after all.

[133] For instance, if it is my volition that you help me get up from a fall I've just taken, and if you volitionally empathize with me in this respect, then you take my volition as an internal reason to yourself have the concordant will that you help me up from the fall. If it is my volition to pick myself up unassisted, and if you volitionally empathize with me in this respect, then you will take my volition as an internal reason to will that you not involve yourself with my effort.

[134] These conditions are equivalent to the more compact conditions I gave in the previous chapter, assuming some very trivial closure of knowledge under known implication.

(b) If *R* wills that *E* to see to it that v, then *E* has a controlling volitional reason to see to it that v.

(2) *R* relies on *E* to see to it that *v,* and accordingly, wills that *E* see to it that *v*, only because (1), namely, *R*'s belief that *E* is trustworthy with respect to *v*

(3) *E* knows that (2)

(4) *R* knows that (3)

(5) If (1) - (3), then *E* has a controlling volitional reason to see to it that *v*

(6) *E* knows that (5)

(7) *R* knows that (6)

These conditions make relationships normative in the following way. The trustee is rationally required to see to it that *v*, which follows from (1), (2), (3), and (5). The trustor knows that he is subject to this requirement, which follows from (4) and (5). Her ability to demand and rebuke the trustee is supported by (6) and (7), as I will explain when I discuss demand. The requirement on the trustee has the form of a duty directed towards the trustor, which, following Hohfeld, can be equivalently characterized as a claim of the trustor *against* the trustee. The trustee's ability to cancel this requirement would then be a power, once again following Hohfeld.[135]

[135] Hohfeld 1923, 36–38. I mean 'duty', 'claim', and 'power' in the Hohfeldian sense, as they have the structure of the "jural relations" that Hohfeld describes under those names. (Strictly speaking, the trustee also has a liability correlated with the trustor's power). It makes sense to speak of the trustee's duty as directed toward the trustor, and thus of trustor's correlative right, for at least two reasons. First, the trustor has *normative control* over the duty on all three counts mentioned by H. L. A Hart: the trustor can cancel the duty before it has been discharged by rescinding her volition; upon betrayal, she can "seek compensation", in a sense that will become clear, as a consequence of the regret that rebukes prompt; and finally, she can also cancel this obligation to compensate her (Hart 184, 1982). The second reason for thinking the trustee's duty is directed towards the trustor is that the trustor is in a unique position to *demand* that trustee do as required

III. Betrayal and Jeopardized Trust

Betrayal occurs when the trustee irrevocably fails to do as trusted *and* when (5) no longer obtains—when in spite of the trustor's continued reliance and belief in his trustworthiness, the trustee ceases to bear the trustor good will. Irrevocable failure—failure to do what was previously possible and is now impossible—is intuitively necessary for betrayal. That good will ceases to obtain is also required because irrevocable failure does not by itself distinguish betrayal from innocent and merely incompetent failure. Consider that we do not normally consider someone to have betrayed our trust whose good faith effort was thwarted by a natural disaster or other *force majeure*. Such a failure is not only not a betrayal, it does not even usually impugn the trustee's trustworthiness, and is in that sense innocent. Additionally, irrevocable failure can be a consequence of incompetence, which does impugn the trustee's trustworthiness, yet still does not constitute betrayal. Consider that there is a very material difference between a prisoner who implicates his accomplice in order to receive a lighter sentence, and one who does so because his simple-mindedness allows the interrogators to trick him into thinking that the accomplice has already implicated both of them. Both failures make the prisoner an untrustworthy partner in crime, but only the former failure seems to qualify as a betrayal.

What if (5) ceases to obtain before the trustee irrevocably fails? This is likely be the case for many betrayals—those that happen not on a lark but with some degree of premeditation. If conditions (1) – (7) are necessary for a trust relationship, however, then most betrayals would not count as such: before they could occur, the trust relationships they presuppose would cease to

(see, e.g., Feinberg 252, 1970). On my account, she is in the position to demand for the same reason she is in a position to control, namely, her leverage over the trustee's internal reasons.

exist. But trust clearly does not work this way, since you do not cease to be able to betray me the moment my rival offers you twenty thousand dollars in exchange for my secret, giving you more internal reason to dish up the dirt than to keep my confidence. Now this is not to say that one cannot prevent a trust relationship and the attendant obligation from arising in the first place: "Before you tell me anything," you might have said as I took you aside, "You should know that I'm not great at keeping secrets. In fact, I hear that someone is offering money for yours—money that I could really use as I'm behind on my mortgage and I don't know what to tell the bank the next time it calls." In this respect, a trust relationship is like a promise: the trustee, like the promisor, can prevent the attendant obligation from arising, but once arisen, he cannot simply cancel it.

In cases of premeditated betrayal, then, (1) – (7) are *initially* satisfied. When the trustee's internal reasons to betray come to outweigh his volitional reason to keep trust, but before the trustor discovers the trustee's change of heart, the trust relationship between the trustor and the trustee becomes a *jeopardized trust relationship*. In a jeopardized trust relationship, (1) – (4) still obtain, but (5) ceases to obtain, and consequently, (6) and (7) cease to obtain. (5) – (7) are replaced by the following weaker conditions:

(5′) *E* has a volitional reason$_R$ to see to it that *v*

(6′) *E* knows that (5′)

(7′) *R* knows that (6′)

(5′) reflects the subjective character of betrayal as something that can be *worth it*—as something undesirable in itself but potentially desirable *on balance*—where the volitional reason to keep trust is outweighed by a conflicting internal reason. Now, it is logically possible that the trustee finds his volitional reason entirely extinguished, in which case it would follow that his

duty to the trustee is annulled. But I cannot think of a psychologically plausible way for this to occur without the trustee's agency being compromised, in which case the annulment of his duty might be a plausible consequence.[136] That said, there *are* cases where trustees retrospectively discover that they did not, in fact, have controlling volitional reasons to do as trusted when the trust relationship was formed. One example would be the mark discovering a confidence trickster's intention to defraud. But in those cases, the trustee did not have a duty to the trustor in the first place, as the trustee had a contrary internal reason *at the time* he acquired his volitional reason, even if it was only later that he discovered it.

IV. Prospective Responses to Betrayal

To best understand why rebuke is a reasonable *retrospective* response to betrayal, it will help to explore the *prospective* responses available to the trustor, as betrayal is deeply linked to one of them. To that end, in this section, I consider the prospective responses to betrayal available to the trustor, including what I take to be the simplest response, namely prejudicial rescission, which has the disadvantage of compromising the relationship between the trustor and the trustee. In contrast, demanding is an alternative response that can avoid this disadvantage, as I will show. I shall then be able to consider how rebuke is illuminated by being an after-the-fact demand.

[136] For instance, the trustee might suffer a head injury and cease to take the trustor's reliance as a reason, becoming a "fiduciary psychopath." I am inclined to hold that fiduciary psychopaths cannot have fiduciary duties, mirroring Gary Watson's position on psychopaths and moral responsibility (Watson 2011).

Prejudicial Rescission

Suppose that the trustor ceases to believe that the trustee is, in fact, trustworthy, raising the prospect of his betrayal. Perhaps she learns that the trustee has a substantial incentive to betray her, and she is not confident whether her volition outweighs the incentive, or vice versa, in his economy of internal reasons. It follows that the first four conditions of both trust and jeopardized trust no longer obtain. This is because the (2), (3), and (4) depend on (1), her belief that the trustee is trustworthy. Consider (2) in particular. Since the trustor no longer believes that the trustee is trustworthy—since (1) no longer obtains—then while she might still rely on the trustee to see to it that *v,* she can no longer do so on account of believing the trustee to be trustworthy. Since, by (2), this belief is necessary for her reliance and the volition that her reliance involves, then assuming that the belief cannot be restored, she is rationally required to rescind her volition, ending her reliance.

Rescission under these circumstances preempts the trustee's betrayal and any harm to the trustor that would follow from it. However, it does so at a cost: by rescinding her volition, she ends her trust relationship with the trustee, annulling the trustee's duty and her own correlative claim. What is more, rescission under these circumstances is *prejudicial* against the trustee: since the trustor doubts that the trustee is trustworthy to see to it that *v*, it behooves her to doubt the trustee's trustworthiness in all relevantly similar circumstances. If I tell you in confidence that I think that a mutual acquaintance of ours is obnoxious, and if I come to doubt that you will keep my confidence, then, other things being equal, it would be irrational for me then to tell you in confidence what I really think about another mutual acquaintance. Now, relevant similarity is not an all-or-nothing affair: I might trust you with less sensitive information—perhaps only with my

neutral or favorable opinions of people—or trust you only when you also have more substantial, non-volitional reasons to keep my trust, such as when I know your secrets, too. Nonetheless, the ultimate consequence of prejudicial rescission must be estrangement, whether by degrees or in one fell stroke, since trust is undeniably a precondition for intimacy.

In this way, prejudicial rescission has much in common with Scanlon's conception of blame. Keeping in mind that for Scanlon, someone is *blameworthy* who reveals "intentions or attitudes that are faulty by the standards of a relationship," he writes:

> To *blame* a person is to judge that person to be blameworthy and, as a consequence, to modify one's understanding of one's relationship with that person (that is, to alter or withhold intentions and expectations that that relationship would normally involve) in the particular ways that the judgment of blameworthiness makes appropriate, given one's relation with the person and the significance for one of what that person has done.[137]

The drawbacks of prejudicial rescission as a response to betrayal, in turn, reflect some of the limits of Scanlonian blame, as well. For example, Susan Wolf has argued that Scanlon's concept of blame does not cover what she calls *angry blame*, which, unlike Scanlonian blame, does not necessarily reflect an impaired relationship, and which is characterized by negative reactive attitudes like "resentment, indignation, guilt, and righteous anger… arising from the belief or impression that the person has behaved badly toward oneself or to a member (or members of a community," attitudes that "tend to give rise to or perhaps even include a desire to scold or punish the person for his bad behavior."[138] Angry blame is, in one respect, the opposite of

[137] Scanlon 2013, 89
[138] Wolf 2011, 337

Scanlonian blame: according to Wolf, "getting angry and expressing it, and demanding a response," in the course of angry blame, "may bring people together and make them closer, rather than pushing them away."[139]

Without speaking to the relative merits of Scanlonian or angry blame, I observe that just as prejudicial rescission is a prospective response reminiscent of Scanlonian blame, there is another prospective response reminiscent of angry blame—reminiscent, at least, in its expressive character and in its tendency to repair relationships. The response I have in mind is fiduciary demand.

Demand

I understand fiduciary demand as a trustor's attempt to shore up her trustee's doubtful good will. More specifically, she attempts to increase the weight of his volitional reason, whose sufficiency she doubts, by acting on his volitional empathy—the same disposition by which he acquired his volitional reason in the first place. To do this, she *manifests* her volition: she does not so much express that she has it, but rather, expresses it in a way that promotes it. For instance, if I say to you, "Remember, I'm trusting you to keep this secret," I am not merely expressing that I have the volition that you keep my secret, and reasoning that you therefore have a volitional reason to do so. Rather, my saying what I said is itself an attempt to promote what I am willing. I am not just telling you of my trust, I am trying to get you to keep it *in* telling you. Manifesting a volition is, in this way, a *trying*, and it is precisely by trying in front of the trustee that she prompts the trustee's volitional empathy in order to shore up his volitional reason.

[139] Wolf 2011, 339

Why, in demanding, must the trustee not only manifest her volition, but know that the trustee knows that her volition is his reason? The trustee, for his part, must know that the trustor's volition is his reason, or he will not receive the trustor's attempt to demand as a demand, but rather, as an inexplicably peremptory first offer of trust.[140] For example, we might imagine that, having forgotten about the paperclips, he calls her from the store to inform her that he is buying the staples and the markers. The trustor then attempts to make a demand, saying, with audible irritation, "Look, I need you to buy paperclips, too. Get the paperclips." If this does not jog the trustee's memory, we would expect him to be taken aback, replying, doubtless to his disadvantage, "What? You want paperclips now?" This reply seems to reject the implication that he *already* had reason to buy the paperclips—that buying the paperclips was *already* normative, since he would be at odds with himself if he didn't do so.[141]The trustor, in turn, must know that the trustee knows, because her demand involves making or implying the assertion that the trustee knows that he has a reason.

What, then, does the trustee do by making a fiduciary demand? The immediate effect is to shore up the trustee's volitional reason to do as he was trusted to do. And in shoring up his volitional reason, she does three things. First, she enforces her claim against him by compelling him to keep her trust on pain of rational incoherence, since by reinforcing the trustee's volitional reason, the trustor has reinforced rational incoherence as an inevitable consequence of betrayal. Second, she avoids being rationally required to prejudicially rescind her volition and being

[140] Or as a demand in a non-normative sense, as one might demand to speak to the manager.
[141] It is in this way that, on my account, the trustor would lack the *standing* to demand, since she lacks the ability to compel the trustee to experience rational incoherence if he does not keep her trust, and in this sense, lacks the ability to compel the trustee to experience keeping her trust as normative. This conceives of standing as a three-place relation involving a claimant, a respondent, and a judge.

thereby estranged from the trustee. This is because she has, at the same time, shored up her wavering belief in the trustee's trustworthiness. Third, she promotes whatever further goal her reliance on the trustee serves, a goal she would otherwise need to give up or promote differently. In this way, we see how it can be reasonable to make a fiduciary demand: other things being equal, it is reasonable for a trustor to seek to enforce her claim, to preserve her relationship with the trustee, and to promote the goal served by relying on him in the first place.

V. Rebuke as a Retrospective Response to Betrayal

We can often express fiduciary rebukes in the same language as fiduciary demands, accounting for the trustee's irrevocable failure with nothing more than a change of tense. Just as "I am trusting you" can express a demand prior to betrayal, "I trusted you" unambiguously expresses a rebuke after betrayal has occurred. On my account, fiduciary rebukes are after-the-fact fiduciary demands in the fullest sense: they are fiduciary demands expressed after the fact of betrayal.

But what of the fact that, for rebukes, betrayal has already occurred? The trustee's volitional reason is the trustor's volition to see to it that *v*, such as my volition that you keep my secret. But it is hard to understand how, for instance, I can still have the volition that you keep a secret you've already disclosed. If I cannot, then there is no volition of mine for you to take as a reason, nor can I shore that volitional reason up by manifesting a volition I do not have. Nor, then, can I enforce my claim, preserve our relationship from estrangement, or further the goal of my reliance—not, at least, in any way resembling how demands accomplish these ends. No hope would remain for explaining the reasonableness of rebuking in the same way as I explained the reasonableness of demanding.

Past-Oriented Volitions

The first step towards understanding rebukes as demands, then, is to make sense of the trustor's volition after the trustee's betrayal, since her volition must survive his betrayal for him to continue to take it as a volitional reason. And he must continue to take it as such for the trustor to shore it up with an after-the-fact demand—with a rebuke. We can understand the trustor's volition by attending to a feature I call *propositional tense,* which is the temporal truth condition of an intentional attitude's propositional content. For instance, the tense of the propositional content of the assertion "Tomorrow I will buy groceries" is the fact that the asserted proposition is true only if the speaker buys groceries the day following the utterance, and without regard to grocery purchases at any other time. *Grammatical* tense reflects the relation between the time of utterance and the propositional tense of what is asserted, denied, feared, desired, willed, and so forth. Accordingly, grammatical tense can change as a speaker's temporal relation changes with respect to a changeless proposition, just as "Tomorrow I will buy groceries" expresses the same proposition today, with the same temporal truth condition, as "Yesterday I bought groceries" expresses the day after tomorrow.

When I confide in you, my volition is that you *not disclose* my secret. The following day, the same enduring volition should have the same propositional content, and accordingly, the same propositional tense. Due to my changing temporal relation to the unchanging propositional tense, the two utterances involve different grammatical tenses. Accordingly, the volition, expressed the following day, would be that you *have not disclosed* my secret *and* that you *not disclose it*. Note that the object of the volition now extends into the past. Suppose that you let news of the party slip, spoiling the surprise. If it is possible for my volition to survive your

betrayal, then grammar poses no obstacle to expressing its content, namely, that you *had not disclosed* my secret. The object of my volition is now not only in the past, but in the counterfactual past.

Can volitions have past objects? This must be possible, and even be common, on the plausible assumption that volitions can endure over time—that something that *was* my will can *still* be my will. Can volitions have objects that their agents believe to be impossible? It might seem that they can't, insofar as one cannot *intend* what one believes to be impossible. However, volitions differ from intentions in this respect. This is because volition, as I understand it, is a conative attitude—an attitude of *trying* to do something—and the ability to try does not seem to require the belief in the probability—or even the possibility—of success. Indeed, several months ago at a local fencing competition, I faced an opponent whom I—as a ticket-bearing spectator—had recently seen representing the United States in an international tournament against the best competitors from around the world, including several Olympic medalists. I was under no illusion whatsoever that it was in my power to beat this opponent. And in that sense, I did not intend to win. But I tried to beat him anyway—it seems strange to characterize my efforts differently—seeing that I fenced him as intelligently and with as much determination as I fenced everyone else, including everyone I would beat, nor did I exert myself any less with each point he scored. Accordingly, if volitions can have both past objects and objects that their subjects believe to be impossible, then the way seems clear for there to be volitions with objects both past and believed impossible—impossible, in particular, on account of being past.

What could a past-oriented volition of this kind be? It would certainly be distressing, considering the conative character of volition and the impossibility of attaining past objects. It

would be an experience familiar to Tantalus, unable to reach the fruit above his head or the water at his knees, both of which would always recede just beyond his grasp. Regret, I have proposed, is one case of this unpleasant attitude.[142] For regret, the subject of the volition and the subject of the volition's content are the same: Williams's lorry driver would will that *he himself* not have run over the child. In this respect, regret is distinct from the volition of the betrayed trustor: she wills that the *trustee* had seen to it that *v*. We might call the trustor's will *volitional blame*.[143] While it is beyond the scope of this chapter to explore the relationship between volitional and existing concepts of blame, I pause to note that volitional blame could explain the relationship between Scanlonian blame and Wolf's angry blame. As we shall see, volitional blame can ground rebuke, in the manner of Wolf's angry blame; as well as retrospective prejudicial rescission, in the manner of Scanlonian blame. An account of volitional blame might in this way explain how angry blame and Scanlonian blame can be, at bottom, blame of the same kind. Importantly, rebuke is not the same as volitional, Scanlonian, or even angry blame, since rebuke is essentially something *expressed,* something which neither Scanlon nor Wolf seem to require.

Rebuke as an Alternative to Prejudicial Rescission

Suppose that the trustor discovers that the trustee has betrayed her. This, naturally, disconfirms her belief in the trustee's trustworthiness. What becomes of the trustor's volition? It is now past-oriented, and she knows it to be impossible to carry out. As I have just discussed, a volition can have these properties, so she must still decide whether to rescind it. Once again,

[142] See chapter 1

[143] The main components of volitional blame would be (i) an agent's past-oriented volition that someone else had done otherwise; (ii) the agent's belief that the blamed party now has and then had reason to have done otherwise; (iii), the agent's belief that the blamed party knew then that he had reason to have done otherwise, and knows now that he has reason to have done otherwise.

assuming that her belief in his trustworthiness cannot be restored, she is rationally required to rescind her volition. Rescission after betrayal is prejudicial for all the same reasons it is prejudicial in the prospective case: the trustor's doubts about the trustee's trustworthiness to see to it that *v* generalize to relevantly similar circumstances. If anything, retrospective rescission is likely to be more prejudicial, insofar as the trustor's belief about the trustee's trustworthiness is more thoroughly disconfirmed by an accomplished betrayal than by evidence of a disposition or settled intention to betray in the future. In this way, rescission still has the effect of estranging the trustee from the trustor.

If a rebuke really is an after-the-fact demand, then it consists in manifesting her volition—now past-oriented—while knowing that the trustee knows that he takes it as an internal reason. But what does this accomplish? Recall that a demand enforces the trustor's claim by compelling him to keep her trust on pain of rational incoherence, it obviates the requirement of prejudicial rescission, and it promotes whatever further goal served by the trustor's reliance. Does a rebuke still promote these ends?

To answer this question, let us consider the effect of shoring up the trustee's volitional reason subsequent to his betrayal. Inconveniently, the trustor's volition is past-oriented, and as a consequence, so is the trustee's concordant will. The trustee is now in the unfortunate position of willing that he had seen to it that *v* while knowing it is impossible now to have done something he didn't. In my first chapter, I interpreted volitions of this form as regret. Intuitively, regret is a good start as far as regaining trust is concerned. Yet we do not typically make demands simply to avoid disconfirming our belief that someone is trustworthy, but rather, to accomplish some further end. I might pay a roommate my share of the electric bill, trusting him to send the entire

amount to the utility company. This I do not for its own sake, but so that the lights work, so that I can charge battery-powered electronics, so that my alarm clock wakes me up, and so on. Accordingly, betrayal usually harms more than one's belief in another's trustworthiness—harm that demands can avert. Can rebukes address the harm of betrayal in addition to the trustee's deficient good will? I submit that they can. This is essentially because the trustee is moved by regret, conceived as past-oriented volition, to mitigate the damage, exactly as we intuitively expect of someone sincerely regretful. We can see how rebukes promote this outcome via regret if we apply two principles concerning volitions that are necessary for an agent's rational coherence.

Two Principles of Rational Coherence

By 'rational coherence', I simply mean the sort of coherence compromised by conflicting beliefs or mutually incompatible intentions—the sort of coherence John Broome considers a *normative requirement*.[144] In Broome's terminology, a normative requirement is a *strict non-detaching relation* of the form $O(P \rightarrow Q)$.[145] $O(P \rightarrow Q)$ is *strict* in that it reflects the 'ought'-like all things considered, rather than *pro tanto,* force of its object. It is *non-detaching* in the sense that O(Q) does not follow from *P*. For example, it is normatively required of me that if I believe that it is Tuesday, then I believe that it is a weekday. Now, suppose I believe that it is Tuesday. It does not follow that I ought to believe that it is a weekday. Rather, the *wide scope* of O means that I ought *either* to believe that it is a weekday, *or* to cease to believe that it is Tuesday, since both

[144] Broome 1999, 411–13.
[145] For Broome, *P* and *Q* are propositions. I have preferred to think in terms of intentional states and events. It should make no difference if one takes Broome's discussion of normative requirements to be more metaphysically neutral, or if one paraphrases everything I say about the normative relationships of volitions and events in terms of propositions.

alternatives would make me consistent with myself. [146] The wide scope of the two principles below will be crucial in the coming discussion.

The first wide-scope rational coherence principle concerns the relationship between volitions and actions. We might call this the *encratic principle*:

> *Encratic principle:* for any agents *A* and *B* and for an event *v*, it ought to be the case for *A* that if *A* knows that she has a controlling internal reason that *B* see to it that *v*, then *A* has the volition to see to it that *B* see to it that *v*.

For instance, if A knows that she has a controlling reason that she herself arrive on time, then she may see to it that she herself sees to it that she herself arrive on time. (The two occurrences of "seeing to it" and of "she herself" are awkward and unnecessary when *A* and *B* are the same agent, but convenient when *A* and *B* are not, as in the case of trust and other reliance relationships.) The encratic principle has wide scope because *A* would be rationally coherent *either* by willing what she has a controlling internal reason to do, *or* by not having a controlling internal reason known to her to do it. It might seem questionable that people are free to simply cease to have a known controlling internal reason, but this seems to be precisely what people are doing when they try to constrain their future selves. For instance, my internal reason to snooze my alarm clock to get a few more minutes of rest might not be controlling if the snooze button triggers a donation from my bank account to causes I strongly oppose.[147]

I call the second wide-scope principle the *volitional coherence principle*:

[146] Schroeder 2004, 337
[147] As of this writing, such alarm clocks are, in fact, for sale.

Volitional coherence principle: for an agent A, it ought to be the case for A that: if A wills that φ and A knows that ψ is the means to φ, then A wills that ψ.

Here I gloss over complications involving multiple means to φ known to *A*. We might assume, for instance, that all things considered, ψ is the best means to φ as far as A knows. The point of this principle is not to capture the nuances of instrumental reasoning, but rather, the intuition that there is something evidently problematic about someone who wills that φ, knows that ψ is *the* means to φ, yet does not will that ψ. For example, if it is really the will of a certain relation of mine not to experience lower back pain, then every morning, he would do the quick and easy stretching exercises that his physical therapist has prescribed, and that he himself grants are effective. There is something deliberatively problematic about not resolving to do the exercises if he was really concerned to alleviate his back pain—problematic, at least, unless he preferred to be in pain so that he can kvetch about it to everyone who will listen.[148]

The normativity of principles like these has been vigorously challenged.[149] For my purposes here, these principles need only be normative in an internal, quasi-psychological sense: one conforms to them on pain of feeling[150] that one has foreseeably undermined one's own ends under conditions of sufficient relevant information. This feeling conforms to considerations of rational coherence. This is because the *fact* of having foreseeably undermined oneself is a straightforward consequence of faulty instrumental reasoning. The *feeling*, in turn, must track this fact given enough relevant information, since the possibility that it does not is precluded as a needlessly uncharitable interpretation of the agent. For instance, suppose that I lose a number of

[148] Fortunately, he resolved this incoherence in favor of resuming his stretching routine.
[149] See Raz 2005, Schroeder 2009, and p. 30n for my view on their criticisms.
[150] By 'feeling', I mean an attitude closely related to regret conceived as a past-oriented volition. This feeling *is* regret when the agent realizes that she has irrevocably undermined a controlling internal reason.

journals, sketchbooks, athletic trophies, and other mementos of childhood in a housefire that my neglect foreseeably permits to happen. Perhaps I repeatedly put off repairing the faulty wiring. After completely accounting for everything lost, I discover that instead of the distress I expected, I feel profound relief—released, perhaps, from the burdens of childhood dreams and parental expectations. Perhaps I do not find that I undermined myself because my feeling does not track fact, even given enough relevant information. The alternative interpretation, which seems far more apt, is that my unexpected feeling is entirely veridical: it turns out that I did not value what I lost nearly as much as I had thought—that preserving it was not, in fact, a controlling reason for me.[151]

Regret Prompts Repair and Repair Prompts Rescission

The trustee's volitional reason compels him to experience regret in that he must, on pain of rational incoherence, will that he had seen to it that *v*. This is an unfortunate situation, seeing that changing the past is clearly impossible. Yet his situation is not hopeless, as on the encratic principle, the requirement he must satisfy has wide scope: he ought *either* to have seen to it that *v, or* to no longer have a controlling volitional reason to have seen to it that *v*. Since he cannot now have seen to it that *v,* his remaining option is to remove his controlling volitional reason.

Can the trustee somehow reduce the weight of his volitional reason so that it is no longer controlling? This is sometimes possible for garden-variety internal reasons arising from desires: by reducing the strength of the desire for a gin and tonic, one might reduce the weight of the resulting internal reason to mix and drink one. One might, for instance, try to conjure disgust by

[151] Perhaps some agents are so incoherent that no such charitable interpretation is available. But the same extreme incoherence would make it hard to impute ends to such agents in the first place.

recalling the occasions where imbibing made one feel terribly ill. Can one likewise reduce the weight of internal reasons arising from volitional empathy? Conceivably, certain people might remind themselves of their highly problematic beliefs about who is a fully-fledged agent, and who is more like a child or even a non-human animal. In this way, they might come to perceive someone's volitions as somehow defective, inferior, or illusory compared with the volitions of a normal agent. I am reminded of Ralph Fiennes's character of SS Commandant Amon Göth from *Schindler's List,* who tells a terrified Jewish girl with whom he is inconveniently infatuated, "I realize that you are not a person in the strictest sense of the word…."

Setting aside that disturbing possibility, the trustee's only escape from regret is to induce the trustor to rescind her volition. Insofar as he values his relationship with her, he would find *prejudicial* rescission undesirable, as it would be estranging. But even if prejudicial rescission were acceptable to him, it is the trustor's alone to grant. By rebuking him, she chooses not to grant it, so any attempt to induce her to grant it anyway is likely to fail.[152]

The option left open to the trustee, then, is to obtain the trustor's *non*-prejudicial rescission. Recall that prejudicial rescission is prejudicial because it results from the trustor ceasing to believe that the trustee is trustworthy. But rescission can be motivated—and, indeed, is

[152] This is not to say that the trustor's prejudicial rescission is impossible for the trustee to obtain on his own initiative. Conceivably, the trustee might convincingly represent himself as incorrigible, though this would likely be false, seeing that if he were really incorrigible, he would be insensible to the trustor's volition, not taking it as an internal reason at all, and not feeling any rational pressure in the first place to have the volition rescinded. Alternatively, the trustee might do the trustor a valuable favor that, though important, does not restore the trustor's opinion of his trustworthiness. In a fictional example from *Pulp Fiction.* Bruce Willis's character Butch, a boxer, is paid by Ving Rhames's character, the gang boss Marsellus, to lose in the fifth round of an upcoming match. Instead, Butch wins the fight, having used Marsellus's money to bet on himself. Marsellus is outraged and resolves to have Butch killed. Eventually, Butch rescues Marsellus from, shall we say, an extremely hazardous situation that they find themselves in together, and in return, Marsellus allows Butch to leave Los Angeles without being pursued.

usually motivated—by non-prejudicial considerations. This is clearly the case with entirely intrapersonal volitions where considerations of trustworthiness do not even arise: on my way to the refrigerator, I might realize that I'm not as hungry as thought I was, and consequently rescind my volition to retrieve a snack. Similar cases occur in trust relationships. For instance, the trustor might rescind her volition that the trustee buy paperclips if she finds a dozen boxes of them hidden at the back of the office supply cabinet. Observe that her rescission under these circumstances has nothing to do with her belief in the trustee's trustworthiness. It is therefore not prejudicial and not tending towards estrangement, as she is not rationally required to modify her other trust relationships with him, or to be more reluctant to enter new ones with him in the future.

Thus, we see that in both the intrapersonal case and in the case of trust, rescission occurs in accordance with the volitional coherence principle, and by taking advantage of the principle's wide scope. Because of its wide scope, an agent ought either to will the means to her end, or to cease to will the end. In my case, since I no longer have the volition to eat a snack, I no longer ought to have the volition to retrieve one from the refrigerator. In the case of the trustor seeking paperclips, since she no longer wills that she remedy the paperclip shortage, she no longer ought to will that the trustee purchase paperclips from the office supply store.[153]

We see, then, that the rational coherence principle allows the trustor to rescind a volition that is a means to an end that she also wills, provided that she rescinds that further volition as well. This is relevant to seeing how a trustee might get a trustor to rescind her volition if her

[153] The placement of *ought* is significant: while I no longer ought, and the trustor no longer ought, it does not follow that I ought not, or that the trustor ought not. In short, the volitional coherence principle no longer requires the trustor and me to retain our volitions, though it permits us to retain them.

volition is a means to a further end that she wills. Is it? I contend that it is, and even that it must be. This is because trust involves reliance, and even without performing a conceptual analysis, it seems fair to say that all reliance, including the reliance of trust, is always reliance *to* some further end: I rely on you to keep my secret, the office manager relies on the intern to purchase paperclips, and so forth. Even reliance on objects is reliance to a further end: boxers rely on gloves to protect their hands,[154] New Yorkers rely on the subway to get to work, and I rely this computer to save this chapter instead of letting it disappear unsaved into the ether. We do sometimes say that we rely on people or on things without further specifying, but it seems clear that in those cases, it is either obvious from context what we rely on them for, or we mean that we rely on them for various and sundry things.

We can now begin to see how regret prompts the trustee to repair the harm to the trustor caused by his betrayal. As we have seen already, given the encratic principle's wide scope, the trustee can still obey the requirement imposed by his volitional reason by inducing the trustor to rescind her reliance volition. Being a reliance volition, it is a means to some further end that she also wills. By the volitional coherence principle, she is constrained to will the means to that end unless she ceases to will the end. Thus, we see that the trustee can facilitate the trustor to rescind her reliance volition by getting her to rescind her own further volition. This may appear to create a regress, but it needn't, provided that the trustor rescinds her further volition because it is *discharged*—because she gets what she wills—and not because she rescinds some still further volition.

[154] Not, apparently, to cushion their blows.

I will not linger on this point, but I take it that a volition is necessarily rescinded when it is satisfied. If it is my will to sit down and I sit down, I may then have the volition to remain seated, and I may esteem the volition that brought about my seatedness as effective at resting my weary legs. But that volition is one I no longer have. This seems to be because volition, as mentioned earlier, is a conative attitude—an attitude of trying. Trying, in turn, seems to presuppose that the object tried for has not yet been attained.

VI. Repair Prompts Rescission

Consider the following case of betrayal. The trustee borrows the trustor's expensive camera for a photography project. While walking down the street, he runs into an acquaintance who marvels at the extravagant purchase he appears to have made. Relishing the mistake, the trustee decides to swing the camera around nonchalantly by its wrist strap, as if he were so wealthy that the cost of replacement was of no consequence to him. "What, this old thing?" he says, as he hurls the heavy gadget around by a flimsy piece of leather clearly not up to the task, even as it occurs to him that the trustor would have never lent him the camera had she known that he would do this. The strap breaks, flinging the camera into the street, shattering the expensive removeable lens and splintering the camera into pieces, which are promptly pulverized under the wheels of a train of school busses returning from a field trip.

Perhaps the trustor regrets his betrayal, or perhaps not at first, not until the trustor rebukes him. This rebuke, manifesting her reliance volition—now her volitional blame—gives him a controlling volitional reason to have a past-oriented volition with an impossible object—to have regret. He cannot obey the requirement in the usual way, by satisfying the volition, namely, to have taken care of the camera. What he *can* do is replace the camera. What replacing it

accomplishes is to satisfy the trustor's further volition, to which her reliance volition was a means. This further volition is to have the use of the camera once the trustee was done with it.[155] By replacing the camera, the trustee satisfies this further volition in an alternative way, by means other than having taken care of the camera. In this way, he induces the trustor to rescind her reliance volition—to let go of her volitional blame—releasing him from the requirement of experiencing regret, and extinguishing his duty to have taken care of the camera.

We can more completely appreciate the way the trustor is induced to rescind her volition by considering both her reason for rescinding and the obstacles to rescission. Regarding her reason for rescinding, recall that regret is unpleasant because it is a volition that cannot be satisfied—a conative attitude towards what cannot be attained. The trustor's enduring reliance volition, now her volitional blame after discovering the trustee's betrayal, is unpleasant for the same reason: as it is now impossible for the trustor to have seen to what he didn't, it is likewise now impossible for the trustor to have seen to it that the trustee have seen to what he didn't. To the extent that volitional blame *is* blame, this is precisely what we would expect. Blame is unpleasant. It takes a toll. And not on the blamed alone, but on the agent blaming. Such is reason enough not to blame, other things being equal.

Unfortunately for the trustor, other things are not equal. While she can rescind her volitional blame, she is constrained in how she can do so, and once again, on pain of rational incoherence. By the volitional coherence principle, volitional blame is required by the further end of her reliance. After all, having seen to it that the trustee saw to it that her camera was taken care

[155] Multiple further volitions are possible. For instance, you might also have the volition to continue to have an asset in exchange for the camera's considerable cost of purchase.

of, is still timelessly a means, albeit now inaccessible, to her further end of having the use of the camera once the trustee was done with it.

With respect to the constraints of rational coherence, the trustor benefits from the same wide scope as the trustee: she can satisfy the requirement of the volitional coherence principle either by having volitional blame, or by rescinding the further volition to which her volitional blame is a means. Now, rescinding the further volition may be very difficult, since she would have to make her peace with not getting the object of this further volition. She would have to make her peace with not having the use of her expensive camera. Reconciling herself to this may, in turn, require that she rescind still further volitions, such as embarking on her own photography project after the camera was to be returned. Alternatively, she might reconfigure her volitions in other ways, such as by rescinding her volition to take a long-anticipated vacation so that she can afford to replace the camera. Such are the harms that, in rescinding her volition without the trustee's repair, the trustor must reconcile herself to suffering. Such are the harms that the trustee repairs when he perfectly navigates the requirements of rational coherence imposed by regret. And such are the harms the trustor compels the trustee to repair by inducing regret with a rebuke.

In the case of the camera, the harm of betrayal can be completely repaired. But sometimes this is not possible. A broken camera can be replaced, though even something like a camera can have irreplaceable properties like sentimental value. What of something less concrete, like the secret of a surprise party? What it if wasn't a camera destroyed or a secret disclosed, but the trustor's child killed in a car accident, the trustee having been provoked into a street race by the sneering driver of a flame-decaled Volkswagen Beetle? In the case of the camera with sentimental value, the harm can still be substantially mitigated, insofar the trustor's reliance is a means to

further ends that don't involve the camera's sentimental value. If the trustee replaces the camera with an identical one, the trustor can still, for instance, embark on the photography project she was planning. In this way, even though the trustee can only mitigate the harm of betrayal, he does make it easier for the trustee to reconfigure her volitions to allow her to rescind her volition of reliance. As for the trustee who discloses the secret of the surprise party, he might make the necessary arrangements to reschedule it so that it's still a surprise. Alternatively, he might lavish his time and resources on the now-expected party so that, surprise or not, it will still be an event to remember for the person to be celebrated, thereby accomplishing a still further end of the trustor's—an end to which preserving the surprise was but a means.

The wide scope of the volitional coherence principle is almost certainly of no use in determining how to repair harms like having killed the trustor's child. The principle can, however, contribute to explaining why this harm is likely irreparable. The trustor willed that the trustee keep her child safe because she valued her child primarily for himself, and not as a means to further ends. Consequently, the trustor has no further end the trustee might satisfy to make it easier for her to rescind her reliance volition.

VII. Conclusion

Using the resources of my accounts of trust and regret, I have sought to show what a fiduciary rebuke does and why it is reasonable. What it does is enforce a trustor's claim against the trustee who betrays her by compelling him, on pain of rational incoherence, to repair both the harm he caused her ends, and the harm he caused their relationship. Both effects are promoted by prompting the trustee's good will, just as a demand does. Prior to betrayal, good will results in the

volition to keep trust. After betrayal, it results in the same volition, namely, to have kept trust—a volition I identified as a case of regret.

The trustee's regret reflects a rational requirement upon the trustee that he repair the harm his betrayal caused to the trustor's further ends. This requirement is nothing more and nothing less than the trustee's original duty to keep the trustor's trust. Owing to the wide scope of this requirement, the duty can be discharged either by keeping the trustor's trust or by inducing the trustor to rescind her volition. Once betrayal has occurred, only the latter option is available. We might find in this situation the beginnings of a duty of repair, seeing that the trustee can induce the trustor to rescind her volition only by repairing the harm to her further ends caused by his betrayal. Often, however, this duty is one the trustee cannot completely discharge. Since the trustor's volition is the ground of the trustee's duty, it remains in trustor's power to rescind the ground and cancel the duty even though the harm she suffered has not been fully repaired. Considerations of rational coherence can make this difficult for her, however, since she would need to reconcile herself to reconfiguring her volitions to account for what is no longer possible, whether that entails exerting herself to repair the harm that the trustee could not, or abandoning the ends placed beyond her reach by the trustee's betrayal.

But even a trustee incapable of repairing the harm to the trustor's ends can repair the harm to her belief that he is trustworthy. For regret itself, however effective or impotent, shores up this belief, since regret is a consequence, and therefore evidence, of the trustee's good will. By shoring up this belief, the trustee removes a cause for the trustor to rescind her reliance with prejudice: that is, to rescind it both in this case and in all other cases where she relies on the trustee for anything relevantly similar. And by removing this cause, the trustee removes a cause

for the estrangement that would follow. This cannot guarantee that the relationship between the trustor and the trustee will be undamaged, or even that it will endure it all. For harm itself is a cause for estrangement, over and above the trustor's doubts about whether the trustee bears her enough good will to be trustworthy.

These, then, are the ways that fiduciary rebukes can allow a trustor to avoid writing off both the trustee himself and the harm he caused. I submit that the latter, on its face, is a reasonable goal. The former, in turn, is reasonable to the extent that the trustor still values her relationship with the trustee in light of the harm he cannot repair. Where she no longer values this relationship, and where he is incapable of further repair, rebukes serve to condemn a him by his own lights—with his own reasons—to a duty he cannot discharge, to a regret he cannot escape. I would not presume to call this reasonable.

Bibliography

"Adolf Hitler's First Anti-Semitic Writing." n.d. Accessed June 21, 2019. https://www.jewishvirtuallibrary.org/adolf-hitler-s-first-anti-semitic-writing.

Baier, Annette. 1986. "Trust and Antitrust." *Ethics* 96 (2):231–60.

Bedau, Hugo Adam. 1978. "Retribution and the Theory of Punishment." *The Journal of Philosophy* 75 (11):601–20.

Broome, John. 1999. "Normative Requirements." *Ratio* 12 (4):398–419.

Cogley, Zac. 2012. "Trust and the Trickster Problem." *Analytic Philosophy* 53 (1):30–47.

Dasgupta, Partha. 2000. "Trust as a Commodity." *Trust: Making and Breaking Cooperative Relations* 4:49–72.

Deigh, John. 1994. "Cognitivism in the Theory of Emotions." *Ethics* 104 (4):824–854.

Feinberg, Joel. 1965. "The Expressive Function of Punishment." *The Monist*, 397–423.

Ferrero, Luca. 2013. "Can I Only Intend My Own Actions?" In *Oxford Studies in Agency and Responsibility: Volume 1*, edited by David Shoemaker, 1 edition, 70–94. Oxford: Oxford University Press.

Flew, Antony. 1954. "The Justification of Punishment." *Philosophy* 29 (111):291–307.

Frankfurt, Harry G. 2008. "Inadvertence and Moral Responsibility." *The Amherst Lecture in Philosophy* 3:1–15.

Gilbert, Margaret. 2006. *A Theory of Political Obligation: Membership, Commitment, and the Bonds of Society*. Oxford University Press on Demand.

———. 2015a. "Acting Together." In *Joint Commitment: How We Make the Social World*, 1 edition, 23–36. Oxford University Press.

———. 2015b. “Acting Together.” In *Joint Commitment: How We Make the Social World*, 23–36. Oxford University Press.

———. 2015c. “Collective Epistemology.” In *Joint Commitment: How We Make the Social World*, 163–80. Oxford University Press.

———. 2015d. “Commands and Their Practical Import.” In *Joint Commitment: How We Make the Social World*, 409–26. Oxford University Press.

———. 2015e. *Joint Commitment: How We Make the Social World*. 1 edition. Oxford University Press.

———. 2015f. “Scanlon on Promissory Obligation: The Problem of Promisees’ Rights.” In *Joint Commitment: How We Make the Social World*, 271–95. Oxford University Press.

———. 2015g. “Three Dogmas About Promising.” In *Joint Commitment: How We Make the Social World*, 296–323. Oxford University Press.

———. 2018. *Rights and Demands: A Foundational Inquiry*. New York, NY: Oxford University Press.

Hampton, Jean. 1984. “The Moral Education Theory of Punishment.” *Philosophy & Public Affairs* 13 (3):208–38.

Hardin, Russell. 2002. *Trust and Trustworthiness*. Russell Sage Foundation.

Harman, Gilbert. 1976. “Practical Reasoning.” *The Review of Metaphysics*, 431–463.

Hart, H. L. A. 2008. *Punishment and Responsibility: Essays in the Philosophy of Law*. 2 edition. Oxford: Oxford University Press.

Helmreich, Jeffrey S. 2011. “Does Sorry Incriminate - Evidence, Harm and the Protection of Apology.” *Cornell Journal of Law and Public Policy* 21:567.

Hieronymi, Pamela. 2008. “The Reasons of Trust.” *Australasian Journal of Philosophy* 86 (2):213–236.

Hohfeld, Wesley Newcomb. 1923. Fundamental Legal Conceptions as Applied in Judicial Reasoning: And Other Legal Essays. Yale University Press.

Holton, Richard. 1994. “Deciding to Trust, Coming to Believe.” *Australasian Journal of Philosophy* 72 (1):63–76.

Jones, Karen. 1996. “Trust as an Affective Attitude.” *Ethics* 107 (1):4–25.

———. 1999. “Second-Hand Moral Knowledge.” *The Journal of Philosophy* 96 (2):55–78. https://doi.org/10.2307/2564672.

———. 2012. “Trustworthiness.” *Ethics* 123 (1):61–85.

Kant, Immanuel. 2017. *The Metaphysics of Morals*. Edited by Lara Denis. Translated by Mary Gregor. 2 edition. New York: Cambridge University Press.

Morris, Herbert. 1981. “A Paternalistic Theory of Punishment.” *American Philosophical Quarterly* 18 (4):263–71.

Nickel, Philip J. 2007. “Trust and Obligation-Ascription.” *Ethical Theory and Moral Practice* 10 (3):309–319.

Raz, Joseph. 2005. “The Myth of Instrumental Rationality.” *J. Ethics & Soc. Phil.* 1:1.

———. 2012. “Agency and Luck.” In *Luck, Value, and Commitment: Themes from the Ethics of Bernard Williams*, 133–61. Oxford, United Kingdom: Oxford University Press USA.

Rosati, Connie S. 2007. “Morality, Agency, and Regret.” In *Moral Psychology.*, 1st edition. Amsterdam: Rodopi.

Roth, Abraham Sesshu. 2004. “Shared Agency and Contralateral Commitments.” *The Philosophical Review* 113 (3):359–410.

Sales, Ben. 2018. “10 Years Ago, the Bernie Madoff Scandal Rocked the American Jewish World. Here’s How Its Victims Have Fared.” *Jewish Telegraphic Agency*, December 20, 2018. https://www.jta.org/2018/12/20/united-states/10-years-ago-the-bernie-madoff-scandal-rocked-the-american-jewish-world-heres-how-those-victims-have-fared.

Scanlon, T. M. 2013. “Interpreting Blame.” In *Blame: Its Nature and Norms*, edited by D. Justin Coates and Neal A. Tognazzini,. Oxford University Press.

Schroeder, Mark. 2004. “The Scope of Instrumental Reason.” *Philosophical Perspectives* 18 (1):337–364.

———. 2009. “Means-End Coherence, Stringency, and Subjective Reasons.” *Philosophical Studies* 143 (2):223–248. https://doi.org/10.1007/s11098-008-9200-x.

Shafer-Landau, Russ. 1991. “Can Punishment Morally Educate?” *Law and Philosophy* 10 (2):189–219. https://doi.org/10.2307/3504911.

Skillen, A. J. 1980. “How to Say Things with Walls.” *Philosophy* 55 (214):509–23.

Stephen, Sir James Fitzjames. 1863. *A General View of the Criminal Law of England*. Macmillan and Company.

Strawson, P. F. 2008. “Freedom and Resentment.” In *Freedom and Resentment and Other Essays*, 1–28. London ; New York: Routledge.

Sussman, David. 2018. “Is Agent-Regret Rational?” *Ethics* 128 (4):788–808.

Tannenbaum, Julie. 2007. “Emotional Expressions of Moral Value.” *Philosophical Studies* 132 (1):43–57. https://doi.org/10.1007/s11098-006-9056-x.

Wallace, R. Jay. 2017. *The View from Here: On Affirmation, Attachment, and the Limits of Regret*. Reprint edition. Oxford University Press.

Williams, Bernard. 1982a. “Internal and External Reasons.” In *Moral Luck*, 101–13. Cambridge: Cambridge University Press.

———. 1982b. “Moral Luck.” In *Moral Luck*, 20–39. Cambridge Cambridgeshire; New York: Cambridge University Press.

———. 1982c. “Persons, Character and Morality.” In *Moral Luck*, 1–19. Cambridge: Cambridge University Press.

Wolf, Susan. 2011. “Blame, Italian Style.” In *Reasons and Recognition: Essays on the Philosophy of T.M. Scanlon*, 332–47. Oxford University Press, USA.